ANOMALY DETECTION IN INDUSTRIAL CONTROL NETWORKS

BY

MUHAMMAD OMER QURESHI

A Thesis Presented to the

DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

COMPUTER ENGINEERING

MAY 2014

ANOMALY DETECTION IN INDUSTRIAL CONTROL NETWORKS

by

MUHAMMAD OMER QURESHI

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

In Partial Fulfillment of the Requirements
for the degree

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

KING FAHD UNIVERSITY
OF PETROLEUM & MINERALS

Dhahran, Saudi Arabia

13TH MAY 2014

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **MUHAMMAD OMER QURESHI** under the direction of his thesis adviser and approved by his thesis committee, has been presented to and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN COMPUTER ENGINEERING.**

Thesis Committee

Dr.Basem Madani (Adviser)

Dr.Zubair Ahmed Baig (Member)

Dr.Abdul-Wahed Abdul-Aziz Saif (Member)

Dr.Basem Madani
Department Chairman

Dr. Salam A. Zummo
Dean of Graduate Studies

25/6/14
Date

عمادة الدراسات العليا
DEANSHIP OF GRADUATE STUDIES

Dedication

Dedicated to my loving and ever supporting parents

ACKNOWLEDGMENTS

All praises belong to Allah and blessing and peace be upon the final Prophet(P.B.U.H). I would like to thanks all those people who helped and encourage me to complete my thesis. First of all, I would like say thanks to my advisor Dr.Basem Madani for his time, guidance and support. His guidance and support made it possible to solve complex problems in this thesis and finish it on time. I would like to thanks the rest of the committee: Dr.Zubair Ahmed Baig and Dr.Abdul-Wahed Abdul-Aziz Saif for their support, constructive criticism and suggestions.

I am also grateful for the Computer Engineering Department and instructors for providing me the support and instilling both knowledge and KFUPM values. I would also like say my thanks to my friends: Haider Ali, Saad A. Khan, Arbab Latif, Saif Ahmed Ghous M. Asim and my manager:Aiman Rashid for their support during my time at KFUPM.

Last but not least, I would like to thanks my wife,sisters,brothers-in-law,nieces and nephew for the cheering me up during difficult times and their support.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

ABSTRACT (ENGLISH) x

ABSTRACT (ARABIC) xi

CHAPTER 1 INTRODUCTION 1

1.1 SCADA Architecture 2

1.1.1 Supervisory Control Layer 2

1.1.2 Process Control Layer 3

1.1.3 Field Instrumentation Layer 4

1.2 SCADA Software 4

1.2.1 Human Machine Interface 5

1.2.2 Logging and Archiving 5

1.2.3 Automation Software 6

1.3 Communication Protocols 6

1.3.1 Modbus 6

1.3.2 Distributed Network Protocol 9

1.4 Motivation and Objective 12

1.5 Research Contribution 14

1.6 Thesis Outline 16

CHAPTER 2 LITERATURE REVIEW **17**

2.1 Vulnerabilities in Industrial Control System 18

2.1.1 ICS Communication Protocol Vulnerabilities 18

2.1.2 Supervisory Layer Vulnerabilities 21

2.1.3 Field layer vulnerabilities 22

2.1.4 Communication Links Vulnerabilities 22

2.1.5 Stuxnet 23

2.2 Related Work 26

2.2.1 Physical Properties Based 26

2.2.2 Conventional IDS 29

2.3 Machine Learning 36

2.3.1 Support Vector Machine 36

2.3.2 K-Nearest Neighbor 36

2.3.3 C4.5 Decision Tree 37

2.3.4 Genetic Algorithm 38

2.3.5 Best First Algorithm 39

2.3.6 The Ripper System 39

CHAPTER 3 DATASET, FEATURE SELECTION AND BASE-LINING **40**

3.1 Introduction 40

3.2 Feature Selection 41

3.3 Dataset 43

3.4 Testing Scenarios 44

3.5 Base-Lining/Profiling 45

CHAPTER 4 ANOMALY DETECTION IN INDUSTRIAL CONTROL SYSTEM USING SUPPORT VECTOR MACHINE **48**

4.1 Introduction 48

4.2 Support Vector Machine 49

4.2.1 Two-Class Classification 49

4.2.2 Multi-Class Classification 52
4.3 Result and Analysis 52

CHAPTER 5 ANOMALY DETECTION IN INDUSTRIAL CONTROL SYSTEM USING K-NEAREST NEIGHBOR 60
5.1 Introduction 60
5.2 k-Nearest Neighbor 61
5.3 Result and Analysis 63

CHAPTER 6 ANOMALY DETECTION IN INDUSTRIAL CONTROL SYSTEM USING C4.5 DECISION TREE 69
6.1 Introduction 69
6.2 Decision Trees 70
6.2.1 Rule Set Classifiers 72
6.3 Result and Analysis 72
6.3.1 Implementation 73

CHAPTER 7 CONCLUSION 79

REFERENCES 80

VITAE 88

LIST OF TABLES

3.1 Normal Operation 47
3.2 Anomalous Operation 47

4.1 SVM 1 Parameter Results 54
4.2 SVM 3 Parameter Results 55
4.3 SVM 5 Parameter Results 57

5.1 kNN 1 Parameter Results 64
5.2 kNN 3 Parameter Results 65
5.3 kNN 5 Parameter Results 67

6.1 C4.5 1 Parameter Results 74
6.2 C4.5 3 Parameter Results 75
6.3 C4.5 5 Parameter Results 77

LIST OF FIGURES

1.1 General SCADA Architecture 4
1.2 MODBUS Communication Messages 9
1.3 DNP 3.0 Network Topologies 10
1.4 DNP 3.0 and EPA model 12

2.1 Man-in-the-middle attack 21
2.2 Communication between PLC and Step 7 through infected .DLL . 25
2.3 Gene Mutation 39

3.1 Average Power Demand 44
3.2 Testing Scenarios. 45

4.1 Support Vector Machine hyperplane for two input class 50
4.2 Average False Positive Rate for SVM 57
4.3 Average Accuracy for SVM 58
4.4 Average Attack Detection Rate for SVM 58
4.5 Average Precision for SVM 59

5.1 Average False Positive Rate for kNN 64
5.2 Average Accuracy for kNN 66
5.3 Average Attack Detection Rate for kNN 67
5.4 Average Precision for kNN 68

6.1 Average False Positive Rate for C4.5 Decision Tree 74
6.2 Average Accuracy for C4.5 Decision Tree 76

6.3 Average Attack Detection Rate for C4.5 Decision Tree 78
6.4 Average Precision for C4.5 Decision Tree 78

THESIS ABSTRACT

NAME: Muhammad Omer Qureshi

TITLE OF STUDY: Anomaly Detection in Industrial Control Networks

MAJOR FIELD: Computer Engineering

DATE OF DEGREE: 13th May 2014

Industrial control Network (ICN) such as Supervisory Control and Data Acquisition (SCADA) system are widely used in industries for monitoring and controlling physical processes. These industries include power generation facilities, oil and gas,telecommunication and transport. The integration of internet exposes these systems to cyber threats. The consequences of compromised ICN are detrimental for a country economic and functional sustainability. In this thesis we are proposing an Anomaly Detection Method for ICN by using the physical properties of the system. We have developed operational baseline of Electricity generation process and reduce the feature set by using feature selection algorithms. The classification is done by using Support Vector Machine, k-Nearest Neighbor and C4.5 Decision Tree. Finally, we present the accuracy results of our proposed anomaly detection method. We have achieved near ideal results for our proposed approach.

ملخص الرسالة

الاسم الكامل: محمد عمر قريشي

عنوان الرسالة: إكتشاف الشذوذ في شبكات التحكم الصناعية

التخصص: هندسة الحاسب الآلي

تاريخ الدرجة العلمية: 13 مايو 2014

أصبحت شبكة التحكم الصناعية (ICN) مثل أنظمة التحكم الاشرافي وتجميع البيانات (SCADA) تستخدم على نطاق واسع في الكثير من الصناعات للمراقبة والتحكم في العمليات الموجودة.

هذه الصناعات تشمل مرافق توليد الطاقة والنفط والغاز والنفايات والمياه والإدارة والاتصالات والنقل. مع دمج خدمة الانترنت يعرض هذه الانظمة للتهديدات والاختراقات.
لذلك فإن العواقب الخطيرة لشبكات التحكم الصناعية تعتبر ضرر أساسي لإقتصاد الدولة واستمراريتها الوظيفية.

في هذه الاطروحة نقترح طريقة لكشف الشذوذ في شبكة التحكم الصناعية بواسطة استخدام الخصائص الفزيائية لهذا النظام. ولقد وضعنا الاساس التشغيلي لعملية توليد الكهرباء وتقليل مجموعة المزايا باستخدام خوارزميات اختيار الميزة.
التصنيف تم باستخدام آلة المتجهات, K - الجار الاقرب وC4.5 شجرة القرار. في النهاية, فإننا نقدم النتائج الدقيقة لطريقتنا المقترحة لكشف الشذوذ. حققنا نتائج اقرب للمثالية للوصول للمنهج المقترح

CHAPTER 1

INTRODUCTION

Industrial Control Systems (ICS) are widely used in industries for monitoring and controlling physical process. These industries include power generation facilities (conventional and nuclear), oil and gas industries, waste and water management, telecommunication and transport. The main motivation for employing ICS for these process is to supervise, control and monitor geographically dispersed resource in these industries. Industrial Control System involves the data transfer between control center, Remote Terminal Unit (RTU)/ Programmable Logic Control (PLC) and operator terminal. This collected data provides information about the states of monitored system. Supervisory Control and Data Acquisition (SCADA) is a type of Industrial Control Systems. It is responsible for monitoring and controlling industrial process to ensure continuous operations and safety of the plant and human beings.

1.1 SCADA Architecture

The SCADA system can be divided into two layers [1, 2], namely client layer and data server layer. The client layer is responsible for Human Machine Interface (HMI), logging of data, archiving and reporting, batch processing while the data server layer handles the communication between RTU/PLC and field devices. However in [3], a three layered open (vendor neutral) SCADA Architecture is presented, they divide the data server layer into, Process Control and Field Instrumentation layer, this three layer architecture has some overlap between process control layer and client layer which is termed as Supervisory control layer in this architecture. Each of these layers is described below.

1.1.1 Supervisory Control Layer

This layer is responsible for providing the Human Machine Interface (HMI) or Man Machine Interface (MMI), periodic data polling Remote Terminal Unit (RTU) or Programmable Logic Control (PLC) devices. The master station is responsible for polling and processing of the data from field devices through Remote Terminal Unit (RTU) and presenting it to human operator in a form that operator can work with [4]. Depending on the complexity of the process control network, the master station can delegate some of its tasks to sub-master station and only received aggregate data from these stations. The Master stations consist of a single computer or multiple networked computers, each with dedicated duties for e.g. archiving, trending, alarm handling etc. The operator station is connected

to the master station through LAN and presented the data on console. It also provide control panel for controlling remote devices.

1.1.2 Process Control Layer

This layer acts as the interface between physical processes and SCADA system. This interface is typically provided through (RTU) or PLC. These are intelligent devices which have control program stored in them which are usually written as ladder logic [3]. These control programs allow changes to running state of the devices attached to them through actuators. Besides controlling the state of the process, these devices are also responsible for polling the data from these devices and transmitting back to Master Station. The transmission of data uses different communication protocol and varies on the type of medium used. PLC is a special purpose computer with memory and I/O, traditionally they dont have communication module but modern PLCs have communication capabilities. Similarly RTUs were generally used to provide I/O interface and communication capabilities and connected to intelligent controller for control program but modern RTU have incorporated the controller, these development blur the differentiating line between PLC and RTU. The I/O interfaces of these devices can be digital or analog; different PLCs come in different configuration. Typically analog I/O are used for measured values for e.g. pressure, temperature or speed while digital I/O are used for presenting the state of a system for e.g. valve state (open/close).

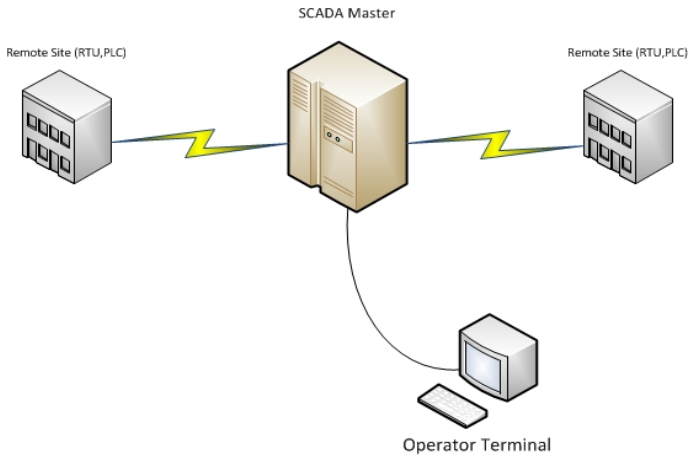

Figure 1.1: General SCADA Architecture

1.1.3 Field Instrumentation Layer

This layer is comprises of different sensors and actuators. The control signals received through RTU are executed through actuators and data acquisition is done through sensors for e.g. reservoir level meters, water flow meters, valve position transmitters etc.

1.2 SCADA Software

The SCADA software can be divided into two categories of open, proprietary and commercial software. The proprietary software are designed for specific hardware and usually main concerned with process control [4], while many modern SCADA systems are using commercial off the shelf (COTS) software for various applications. The main objectives of COTS are to be compatible with wide range of SCADA products. The main functionalities of the SCADA software are discussed below.

1.2.1 Human Machine Interface

This software provides the interface to an operator to interact with the controlled system and provides the holistic view of controlled system for e.g. Power generation plant by integrating data from different sources through Master Station. HMI software is available as commercial software for e.g. Systemview [5] and usually installed on Microsoft Windows platform. The HMI provides a holistic view of the controlled process to an operator by integrating data from different RTU through master station. Modern HMI software provides rich graphical visualization of the controlled process. The other functionality of HMI software includes alarm management; it displays alarms from different devices to an operator and provides control options to manage these alarms.

1.2.2 Logging and Archiving

A scalable and robust logging system is needed for a SCADA system because the amount of information from RTUs is massive. The need for logging the data is essential for a SCADA system, as it is need to retrieve the information about device operation statistics. The logging is done for the most recent data, while archiving is storing the logged data for long term use. The logging frequency is dependent on the nature of the application; it is performed on a cyclic basis or can be event triggered. The logged information from logging or archive database is used to perform trend & analytical analysis. The logging system integrates with HMI and trending systems.

1.2.3 Automation Software

Automation software is used to configure and maintain control application in RTU and PLCs. Each vendor has its own software that configures its PLCs or RTU for example Siemens PLC are configured through Step 7 [6] software by using ladder logic as option. The automation software configures the PLCs / RTU control logic and defines action on different events.

1.3 Communication Protocols

The communication between RTU / PLC and Master station is governed by set of rules and standard, this include addressing scheme, message format and data types. Master station send command to RTUs to perform an action or request for information and RTUs respond to these commands, the rules for each type of such transaction is set by communication protocol. There are many protocols in use but the most popular protocols are Modbus, Distributed Network Protocol 3.0 (DNP 3.0) and IEC 60870-5-101.

1.3.1 Modbus

MODBUS is an application layer messaging protocol, which establishes rules and message structure for the communication between endpoints (PLCs) and Master Station. It operates on the principle request/reply mechanism and defines services by different function codes, where masters station send request for service by specifying appropriate function code to slave devices and responses from slave

devices and use function code to represent the outcome of the performed action. It was developed in 1979 by Modicon as a serial line protocol, which is currently owned by Schiender Electric. MODBUS protocol define application layer, which assume abstract transport layer and allow the exchange of the data, independent of the underlying communication layer [7]. This allow the MODBUS to be deployed over Transmission Control Protocol (TCP), currently there are two variants of MODBUS

i Modbus Serial

ii Modbus TCP.

Modbus over Serial Line

MODBUS Serial Line protocol is a Master-Slave protocol [8], and operates at data link layer of Open Systems Interconnection (OSI) model. In this model only one master (at same time) is connected to maximum of 247 slave devices via shared bus, the master issues commands to slave for an action or information and receive responses in the results of these commands. Slave devices dont communicate with each other and dont initiate the communication. The Master can issue commands to slave device in either unicast mode or broadcast mode. At the physical layer, the MODBUS can employ different physical layer protocol for e.g. (RS485, RS232). TIA/EIA-485 (RS485) [8].

Modbus over TCP/IP

The MODBUS over TCP/IP is based on client/server architecture and provides interconnection between different devices in a MODBUS network and also between different MODBUS networks. The MODBUS over TCP/IP is implemented by encapsulation MODBUS data in a TCP segment, TCP port 502 is assigned to MODBUS protocol. In TCP MODBUS, slave devices listen for incoming connection on port 502 and act as servers, the master initiate a connection by generating Application Data Unit (ADU), encapsulate it in TCP segment and send it over the network. In TCP MODBUS a master can have parallel communication or transaction; similarly a slave can also communicate with multiple master stations [9]. The semantic of ADU and addressing is discussed later in this section. The TCP/MODBUS defines four types of messages for communication purpose [9],

i MODBUS Request: The request initiated by Master to ask slave to perform an action or request for information.

ii MODBUS Indication: The request message received by slave device or server.

iii MODBUS Response: The message generated by server in response of MODBUS request by client

iv MODBUS Confirmation: The response message received by client.

The communication with above mentioned messages between Modbus client and Modbus server is show in 1.3.1.

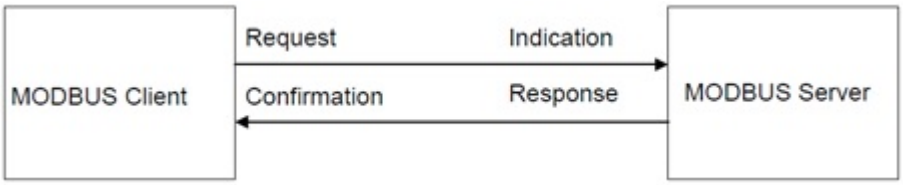

Figure 1.2: MODBUS Communication Messages

The advantages of MODBUS over TCP/IP as compare to MODBUS over serial include economic benefits and interpretability benefits. The TCP/IP networks are ubiquitous, it save cost of dedicated point to point communication link or leased lines, it also provide integration with different COTS application on the existing network and any new device with TCP/IP stack installed can be connected to TCP MODBUS network.

1.3.2 Distributed Network Protocol

DNP was originally created by Westronic, Inc. (now GE Harris) in 1990; it sets out the rules for exchange of control commands and data between devices in SCADA control system [10]. This protocol has been widely used in electric and water companies [11]. DNP is a layered protocol but instead of adhering to seven layers OSI model, it operates on simplified 3 layer standard proposed by the IEC (International Electrotechnical Commission), which is known as Enhanced Performance Architecture (EPA). The DNP3 can be implemented in various network topologies which can be simple direct connection between master and outstation or hierarchical design for complex networks. Three common network topologies are shown in figure below. The top network topology is simple one on one con-

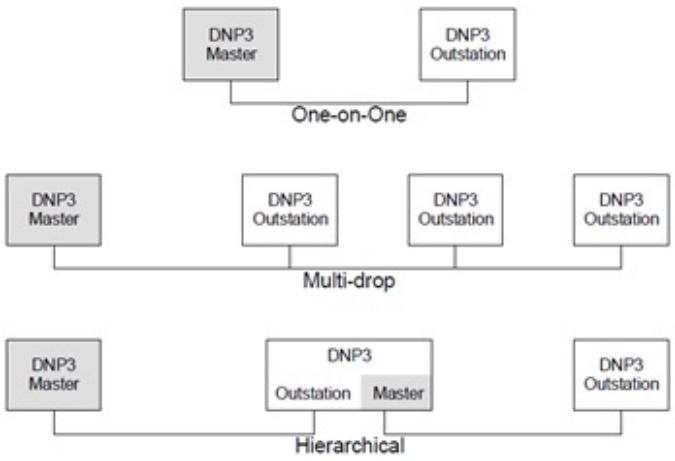

Figure 1.3: DNP 3.0 Network Topologies

nection between master and outstation, the connection between them can be a dedicated line or dial up telephone line [11] [12]. In multi-drop design several outstation is connected to a master station, it queries data from outstation in round robin order. Every outstation listens to every request from master but only permitted to respond to request that are addressed to it. The physical connection between master and outstation is multi-dropped phone line, fiber optic or radio [11]. In hierarchical device, a device acts as an outstation in one segment and masters in another segment, this device also act as data concentrators or protocol converter.

EPA model is consists of Application, Data link and Physical layer; however DNP adds a fourth layer, pseudo transport layer for message segmentation. Each layer function and semantics is discussed below.

Physical Layer

The physical layer is responsible for managing physical media resources by monitoring its states, maintaining synchronization, controlling voltage and etc. DNP can be transported over different kind of media including fiber, copper, radio or satellite. Recently DNP is also implemented specified over simplified serial physical layer using fiber, copper, radio or satellite. DNP can also be transported over Ethernet or TCP/IP .

Data Link Layer

The data link layer manages the logical link between communicating devices, ensuring the reliability by improving physical channel error characteristics . DNP3 adds a 10 byte header in every data link frame and 16 bit error checking for every 16 byte of the frame , the maximum size of data content is 250 bytes. Thus maximum length of data link frame is 292 bytes.

Pseudo Transport Layer

This layer is responsible for managing message fragmentation and reassembly. It is necessary to split application messages larger than one data link frame length into multiple frames. To facilitate the reassembly and fragmentation, one byte header is added to every frame. This byte is consists of three fields FIR,FIN and Sequence number. The FIR and FIN fields indicate the first and final frames of the indicated message, while the sequence keeps tracks the fragment, its incremented for each successive frame.

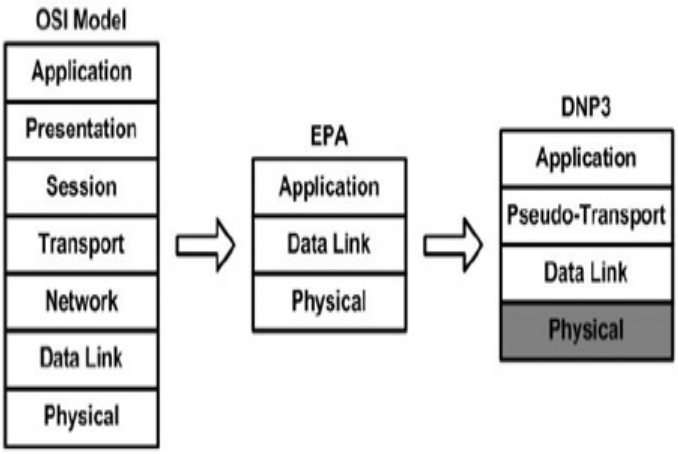

Figure 1.4: DNP 3.0 and EPA model

Application Layer

The application layer acts on the message received and respond with requested data, it defines the roles of the master and outstation devices. The request messages are sent by master devices while an outstation can send solicited or unsolicited messages. Typical master devices messages include request for performing a task or request for data, an outstation message are reply to request, ACK, NACK for acknowledgement purposes. It is possible that application layer fragment messages that exceed the maximum allowed message fragment size (between 2048 and 4096 bytes).

1.4 Motivation and Objective

The introduction of information and communication technology (ICT) into Industrial Control System (ICS) enhances their functionality. The enhancements include better supervision and management of these networks. However, this en-

hancement comes at the cost of the security vulnerabilities and threats. These vulnerabilities are beside the inherent vulnerabilities in ICS e.g. vulnerable protocols. The threats against ICS are legitimate, several attacks against these systems were reported which compromise them. These attacks include Denial of Service (DoS) and reconnaissance attacks. A brief description of such attacks is as follows,

i On Nov 8 2011 [13], a pump at central Illinois water system was shut down remotely. The hackers gain remote access to the network of a water utility with stolen credentials. They managed to shut down one water pump at the facility. The forensic investigation has shown that a computer was hacked from Russia and this attack damages the water pump [14]. This attack exploits the hardcoded password vulnerability in the software of control system.

ii A reconnaissance attack was launched on US electricity grid [15]. After gaining access to the system, a back door and malicious software was installed. The installed software had capabilities to inflict damage on the control system. In 2008, an attack on electric left several power generation equipment inoperable.

iii In January 2003, a nuclear power plant in Ohio, United State. An infection disabled the safety system of a nuclear facility. The worm entered to the nuclear plant corporate network through unsecure network of the contractor. It infected one unpatched windows server in the plant network. Slammer reportedly also affected communications on the control networks of at least

five other utilities. It floods the network which forces the system to drop legitimate control system traffic.

Beside the threats of the attack, human error is also one of the factors that result in the undesired operation of the system. The operators of such systems have to monitors numerous readings and alarms. They take corrective action when an alarm is raised. Most of the approaches and techniques to secure the industrial control system are adapted from data network security. These systems are fundamentally different from data network systems. The security approaches and techniques include signature based intrusion detection system, anomaly detection system using normal communication patterns, command and response patterns and resource utilization profiles. These techniques detect the anomalous operation or attack after it occurred. The aim of this work is to propose an Anomaly detection technique which will use physical properties of the system to detect anomalous behavior of the system due to malicious behavior or fault, while incurring minimum overhead. The proposed system detects the anomalous behavior of a system before it enters into a dangerous state.

1.5 Research Contribution

The main contributions of this thesis are:

i. Dataset Modeling for Power Generation Process: One of the main problems in research on security of the industrial control system is lacking of the real dataset. The industries and industrial control system vendors are reluctant

to share data about their systems. The reason behind this is to keep their technologies and industrial process information confidential and out of hand of hackers. In this research we developed a realistic dataset of the power generation process based on the real data. We model the attack on the power generation process through establishing the normal profile of the process. An anomaly is defined as deviation from the normal profile.

ii. Detection of Attacks in ICS Using SVM: We use the support vector machine to create the classification of the power generation operational data to distinguish between normal and anomalous operation of the power generation plant. SVM are known for higher accuracy and less prone to over fitting.

iii. Detection of Attacks in ICS Using kNN: We propose k nearest neighbor algorithm based anomaly detection system for industrial control networks. It models the normal operation of the plant and detects anomalous operation by comparing the future operational data with model. The proposed approach is subsequently tested with different value of k; analysis of its impact on accuracy, false positive and false negative is also performed.

iv. Detection of Attacks in ICS Using C4.5: We apply decision tree based approach to create a classification of the power generation data that discriminate between normal and anomalous operation of the power generation. C.45 algorithm can be applied to both categorical and continuous data and robust to noise.

1.6 Thesis Outline

In chapter 2, extensive literature review is done on vulnerabilities in industrial control system and related work on anomaly detection in industrial control systems. We also review machine learning techniques used in anomaly and intrusion detection systems.

In chapter 3, the dataset of power generation plant, description of the parameters and feature selection using genetic algorithm is discussed. The normal and attack definition is also established by base-lining of the selected features.

In chapter 4, 5 and 6, we study Support Vector Machine, k-Nearest Neighbor and C4.5 respectively for anomaly detection in the industrial control system. The performance analysis of these algorithms is evaluated by using attack detection rate, accuracy and false positive rate.

CHAPTER 2

LITERATURE REVIEW

The security of the Industrial Control Systems is critical for plant operations and safety. Any disruptions in the services monitored and supervised through ICS have adverse consequences. Typical services include power generation and distribution, water treatment plant, nuclear facilities. Traditionally, SCADA system are believed to be secure from external threats as they are not linked to external networks and used proprietary protocols. The introduction of Internet and commercial applications has numerous benefits but also expose them to security threats Furthermore the most prevalent SCADA communication protocols have inherent security weakness due to lack of authentication, integrity check and encryption. These weaknesses can be exploited to gain unauthorized access, overwrite running configuration, and execute denial of service (DoS) and man-in-middle attack. In this chapter extensive literature review is done on vulnerabilities of industrial control systems and security approach to secure industrial control systems.

2.1 Vulnerabilities in Industrial Control System

The integration of the latest information and communication technologies into the industrial control system enhances its functionalities and connectivity. Once isolated industrial control system are now tightly integrated with internet and use the same network used by other TCP/IP devices. This integration exposes these systems to all vulnerabilities of these technologies. ICS system used to have proprietary software but current ICS uses commercial software with known vulnerabilities for e.g. Microsoft Windows, SQL etc. Beside these vulnerabilities, communication protocol and infrastructure of ICS lack basic security mechanism such as authentication and integrity checks. These vulnerabilities expose ICS to multitude of threats, which can result in catastrophic events. In this chapter we discuss the security vulnerabilities in ICS infrastructure and communication protocols and their impact if exploited. A literature survey of security measure and approaches to secure these systems and brief overview of the machine learning techniques used in our work is provided.

2.1.1 ICS Communication Protocol Vulnerabilities

The communication protocols of Industrial control system are designed to ensure availability, reliability and timely execution of tasks. The majority of industrial control communication protocol are designed and developed when security was not the prime concern. This results in the inherent security vulnerabilities in these protocols. Basic security features e.g. authentication, integrity check and

time checks are missing in these protocols. These vulnerabilities if exploited can results in the compromise of the system and providing complete control to the attacker.

In [16][17], the vulnerabilities of two major communication protocol MODBUS and DNP 3.0 of industrial control system are presented. Although the vulnerabilities are specific to these protocols but they are common among other communication protocols. The potential impact of these vulnerabilities, if exploited is discussed below.

Lack of Authentication

The absence of the authentication mechanism can be exploited to gain unauthorized access of the system. This absence of authentication between master and slave device can be exploited to send malicious command [18] [17]. These commands will be executed by the slave device which can results in disruption of service. This attack is demonstrated in [19], by sending the malicious command, which instruct the device to reset the counters. An implanted device can be used to launch a denial of service attack on industrial control network. Due to the lack authentication, this device can send meaningless data continuously over the communication link [18]. This effectively saturates the communication link and communication between master and slave device will be lost. Another way to launch a DoS attack is to continuously send commands to a particular device or node. The processing and execution of each command consume memory and processing power of a device, continuous command execution will exhaust it

resources. Generally these are resource constrained devices.

Lack of Confidentiality

The communication between devices in the industrial control system is done without encryption. This vulnerability can be exploited to launch the reconnaissance attack on the industrial control system. In this attack, an attacker can capture the communication between devices on the network and can learn about the devices architecture of the network. In [16] [19], this vulnerability is exploited to capture data from the network. They were able to capture the messages between devices. A typical message contains the field device address, response/request and its associated data.

Lack of Integrity Check

The lack of confidentiality can be also be used to launch the man-in-middle (M-i-M) attacks. A message can be capture, modified and retransmit to the addressed device. Due to lack in integrity e.g. CRC, this modification is undetectable. This kind of attack can be used to deceive the security mechanism where only trusted device can send instruction to slave devices. This attack can be used to modify a legitimate command to issue a malicious command. A successful M-i-M launched by exploiting this vulnerability on a test bed [19].

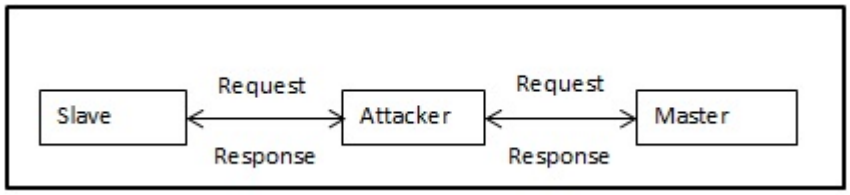

Figure 2.1: Man-in-the-middle attack

Replay Vulnerability

If the communication messages are not time stamped and does not have an expiry limit, such protocols are susceptible to replay attack. The messages can be captured and stored. Such message can be retransmitted at later time to the devices. This vulnerability can also be exploited to launch man in middle attack.

2.1.2 Supervisory Layer Vulnerabilities

An operator system is used to monitor and supervise industrial process. A compromised operator station can be proved lethal. This is one of entry point for an attacker to launch an attack. Following are possible threats and vulnerabilities present in an operator system. 1. The operator workstation have evolved into PCs [20] and based on Windows Platform. It is known that Windows platform have high number of known vulnerabilities and unpatched PC can be easily exploited. 2. The system can also be compromise due to weak Passwords, installation of unauthorized software, connection to internet and absence or obsolete Anti-Virus. 3. The use of commodity hardware and software solution such as Microsoft Windows, TCP/IP networking, SQL database are few example [21], this adoption is mainly due to their low cost, high availability and high connectivity requirement.

This exposes control systems to same vulnerabilities of these products, which can be exploited to gain remote access of the system or perform illegal activities.

2.1.3 Field layer vulnerabilities

RTU are responsible for managing field devices, Programmable Logic Control is commonly used to control filed devices and gather telemetric data. PLCs are programmed by using software, for example STEP 7 by Siemens. The potential for reprogramming an RTU or PLC by accessing the polling/communication circuit exists. The attacker can replace the valid configuration file by malicious file [22]. This threat could be exploited in case of PLC dont provide means of authentication or have hardcoded username and passwords in the firmware.

2.1.4 Communication Links Vulnerabilities

Communication links are integral part of Industrial Control Systems and provide connectivity between remote sites. The remote sites can be connected through dial-up link, leased lines, frame relay, wireless links, and wired links or through internet. Following are the vulnerabilities or threats which can lead to intrusion into an industrial control system network.

i War Dialing : Vendors and support staff use dial-in functionality to gain administrative access to Industrial networks. They establish by using dial-up functionality and authenticate by username and password [21]. War dialing can be executed using scripts to dial surrounding numbers to establish a

connection. The default username/password of the devices can be used, which is set by the vendor of the device [22].

ii The communication links connecting remote sites are usually wireless links. The wireless links are vulnerable to these communications hijacking [21]. The communication link between RTU and MTU can be hijacked by doing the man in the middle attack. This attack could result in sending false information to an operator. Any action based on this information could do potential damage or interruption of process.

iii The sites which are connected through internet are more vulnerable and highly susceptible to attacks. The TCP/IP protocol does not have any inherent security mechanism against attack. If an adversary is able to gain access of one computer/node, it can easily traverse through network. Worm and viruses also exploit this functionality. The communication links are also susceptible to electronic intrusion, signal jamming, RF eavesdropping, sneaker net and etc.

2.1.5 Stuxnet

Stuxnet is a complex malware that target Industrial Control System networks. It is specifically created for targeting Industrial Control Systems with the final goal to reprogram the Programmable Logic Control (PLC) devices; which are used to control the physical processes. Stuxnet is a highly targeted malware and only infect specific types of PLCs, precisely Siemens PLCs with 315-2 CPU.

STUXNET Architecture

Stuxnet has complex architecture and consists of different components. Mainly it is a large .DLL file which contains different exports and resources. Each export and resource has a distinct function in controlling and execution of the Stuxnet. Exports are used to perform different tasks e.g. contacting command and control center or infecting removable drives while resources is consist of drivers, exploits, .DLL files, cabinet files and PLC root kit. Exports use these resources to control the operation of the Stuxnet. Beside exports and resource, Stuxnet also contain two encrypted configuration block. This configuration data instruct the Stuxnet how to act on a comprised system.

STUXNET Infection and Installation

Initially STUXNET spreads through removable media and copying itself over the network. It exploits two zero day vulnerabilities for network propagation. These zero day vulnerabilities are related to printer spooler service and windows server service. Once Stuxnet malware is dropped, its installation process starts. The zero day exploits were used to escalate the privilege level to initiate the infection process. The antivirus and intrusion detection software were unable to detect this infection. The process for injection will be chosen according to security software e.g. winlogon.exe for McAfee. Stuxnet will use this process to call the export 16; responsible for Stuxnet installation. Once the infection process is completed, it replaces the .DLL file of PLC programming software.

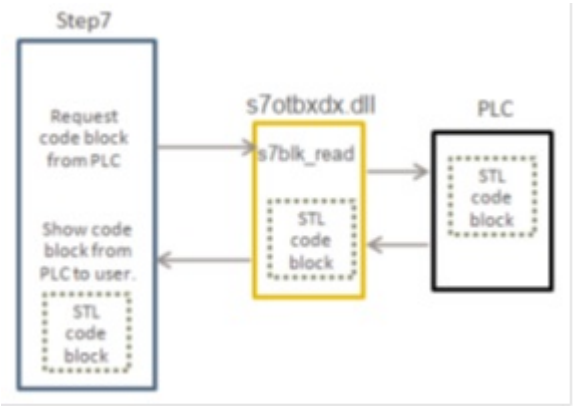

Figure 2.2: Communication between PLC and Step 7 through infected .DLL

The infected .DLL enables Stuxnet to monitor the exchange between PLC and programming device. It can also inject the malicious control program and hide it. The Communication between PLC and Step 7 through infected .DLL is shown in figure 2.2.

PLC Modification

The modification in the PLC control program is decided according to the devices connected to it. Stuxnet has three control program configurations named A, B, and C. Stuxnet only infects PLCs with 6ES7-315-2 CPU. The infected .DLL injects the malicious code into the PLC to overwrite the current configuration.

Equipment Damage

The malicious code, damages the variable frequency converter drives attached to PLCs. The infected sequence damages the drives by slowing or speeding up them at different frequency which lies outside their normal operation range. The malicious code changes the operating frequency periodically. The normal operating

frequency between 807 Hz and 1210 Hz.

2.2 Related Work

The security of SCADA system can be achieved by two approaches: Securing the perimeter of SCADA network by firewalls, Intrusion Detection System or Anti-virus, the second method is to develop profiles of normal operation and detect intrusions using these profiles [23]. The second approach can be further divided on the basis of what feature sets are used in developing these normal operation profiles. The feature sets can be related to protocol parameters, network traffic pattern or measurements from physical process for e.g. pressure, speed, power etc. The following section provides a detail survey on the research which uses the physical properties of the control system. A survey about IDS or Anomaly Detection System that use features like protocol parameter or network traffic pattern are provided in the later section.

2.2.1 Physical Properties Based

Alavaro et al. [20] present their finding about incorporating physical system knowledge enable them to identify critical sensors and attacks on them. They modeled Tennessee-Eastman process control system (TE-PCS), which is responsible for controlling a chemical reaction with primary objective is to maintain pressure around 3000 KPa. Three types of attacks, namely surge, bias and geometric are modeled and launched on different sensors in the network. The detection

method use by Alvaro et al. [20] is based on change detection by Cumulative sum (CUSUM), which detect change between two Hypothesis (H0 and H1) in minimum possible time. They find that DoS attacks dont force the control system to operate in unsafe condition but however attacks on integrity of sensor do force the change in pressure beyond safety limits. In [20], an automatic detection module (ADM) is also proposed which replace sensor measurement by estimated measurement by using linear model of system when an intrusion is detected.

In [24], a neural network based intrusion detection system (IDS) is proposed for SCADA lab system for water tank storage. Their proposed IDS used four features as input which includes parameters which presents physical states of the system; water level, mode of operation and water tank pump state. Their attack model consist of command and response injection into the SCADA network which affects the integrity and DoS attacks which affects the availability aspect of the SCADA network. They are able to achieve high accuracy for detecting command & response injection and DoS attacks but have high number of false positive in detecting replay attacks. The proposed IDS approach is designed for water storage tank with a simple process to monitor we suspect that if the same approach is applied to more complex system, the IDS accuracy may degraded.

Bigham J. et al. [23] proposed two approaches to enhance the accuracy of anomaly detection in SCADA systems. Their first approach is based on N-gram technique, which first records the normal operation of the SCADA network The typical use of N-gram is to classify the text independently of language or error,

but they used it model the normal operation of the electric network. It is done by recording n-gram occurrence in a training set. The second approach is based on the invariant induction which establishes mathematical relationship between different data readings and uses these relationships to model the normal operation of the electric network. It detects the corrupted measurement due to fault or attack by evaluating it against mathematical model. They test the performance of their purposed approach they used load flow program to record real and reactive power flow measurement of six bus electric network of varying load for one year period. They introduce 1 to 44 random errors in the calculated data files. In [23], they concluded that invariant induction has better overall performance, while n-gram is better in detecting corrupt files. It is also suggested to use correlation of two or more anomaly detection techniques for further enhanced accuracy.

In [25], they extended the work proposed in [23] by using Bayesian based correlation that correlate the output of two IDS. They used invariant induction and artificial ant approach based IDS. They defined three invariant checkers for profiling and anomaly detection. A similar approach is used for artificial ant approach is adopted by clustering real and reactive powers to define normal operation of electric network.

An ADS technique based on data rough set theory [26] is proposed for securing the Electric Power System. A similar approach to [23] [27] is taken to profile the normal operation by extracting rules from the normal operation and compare incoming measurement from RTU to this normal profile. They have reduced the

number of rules for anomaly detection which makes the rule dynamic and less resource intensive. The experimentation is done on the six bus power system, the data set is consist of reading of 45 test with 57 measurement values. The errors are introduced on bus 4 and bus 6; they only introduce switch sign error in their dataset. The rules for detection anomalies are based on the power flow between the buses on the electrical network. After doing the literature survey for anomaly or intrusion detection technique for electrical networks based on physical properties, we found out that prime focus is detecting anomaly or intrusion from generation output to distribution network and dont taking account the possibility of attack at the machine side of the generating equipment.

2.2.2 Conventional IDS

Intrusion Detection techniques can be classified into signature based and anomaly based. Signature based IDS analyze the network traffic against rules in their databases, this require the prior knowledge of the exploits and vulnerabilities to develop signature, due to this some unknown attacks can go undetected. The level of sophistication of detection techniques in signature based IDS vary and could be simple analysis of the particular fields in the network traffic to deep packet inspection. Popular signature based IDS include Snort [28] and Bro [29]. On the other hand anomaly based techniques detects suspicious events that dont match the normal profile of the system, this helps in the detection of the unknown attacks. The challenges in the anomaly based detection high number of false positive if the

accurate profiling of normal operation is not done. In this section we discussed research for IDS for control networks that use signatures, profiling or model based approaches to detect intrusion attempts or attack on control networks.

Anomaly Based

In [30], an IDS is proposed that modeled the specification of the MODBUS/TCP; they exploit the static nature of control system network topology and network traffic pattern [30][31] . They proposed three techniques to detect intrusions on the control network. The first approach is based on the protocol based specification of the fields in the request and reply messages of MODBUS/TCP. They have developed specification for function codes, exception codes, protocol identifiers and cross field relationship specification for e.g. relationship between length and function code. The second approach exploits the predictable communication pattern in control systems; Snort [28] is used to detect suspicious communication patter. The snort rules are developed based on the modeled communication pattern in MODBUS/TCP based network. In the third approach they used heuristics to learn the availability of the server and clients on the system; this helps them in identifying rogues devices and changing network services. An appliance based on these techniques is integrated into control system test bed, initial experiment produce promising results and provide evidence that model based approach can be used to effectively detection anomalous behavior in control systems.

As mentioned in [30], inaccurate model of the system produce high number of false positives, an accurate modeling for MODBUS/TCP protocol is proposed in

[31] and an IDS is designed to detect attacks. They have used Moor Deterministic Finite Automata with four tuple features; function code, Response/ Query indication, reference number and count, to model normal operation between HMI and PLC. Their proposed IDS performed deep packet inspection and produce detail network profile. They did not inject any kind of malicious packets into their system and considered faults due to misconfiguration and troubleshooting, under this condition the experimental results shows that the false positive rate is very low. They have claimed of zero false alarms for monitoring MODBUS/TCP network for 111 hours.

Yang et al [32] proposed an Anomaly based IDS for control system. Their proposed system is based on Auto-Associative Kernel Regression (AAKR) model and Statistical Probability Ratio test (SPRT). They have simulated the SCADA system by establishing a local network of SUN servers and use SNMP to collect the data from devices on the local network; the monitored parameters include link utilization, CPU usage, and login failure. They used Continuous System Telemetry Harness (CSTH), to monitor the network and establish the base profile of the normal working condition. This base line profiles are classified by time of day, day of week and special days such as weekend and holidays. They used the monitored feature as input to Auto-Associative Kernel Regression (AAKR) model to predict the normal behavior; the incoming data is then compared with this predicted normal behavior. The residual of this comparison is an indicative of the abnormal behavior. These residual are than fed into Statistical Probability

Ratio test (SPRT), this determine whether this deviation from normal traffic is due to normal or abnormal distribution by comparing it to predetermined limits of the parameters. They have considered DoS attacks, ping ood, jolt2 attacks, bubonic attacks, simultaneous jolt2 and bubonic attacks to demonstrate the working of the proposed Intrusion Detection System.

Another Anomaly based Intrusion Detection system is proposed in [33], their approach is based on bloom filters. They considered the typical master-slave topology of the control system with MODBUS as the communication protocol and considered only external threats which are primarily attacks on the integrity of the system. The threat model includes man in the middle attack and compromise of the operator station with HMI to send disruptive commands to devices. They have considered function code and data as the parameter for establishing normal profile. They use n-gram analysis to extract these parameters from the traffic; the extracted information is act as an input to bloom filters. Each bloom filter is assigned with weight, in their implementation they have assigned equal weight 0.5 to each bloom filter, cumulative score of these two bloom filters is used for anomaly detection and decision. An enhancement to their approach is done by adding physical operating mode of the plant to anomaly decision making process. The physical operating condition can be either of normal, emergency and restorative. They used these states to detect anomalous behavior of devices in particular state condition for example repeated restarting of devices in normal state is anomalous behavior while it could be seen as normal behavior in restorative mode.

An anomaly based IDS is proposed in [34] which detect intrusion by establishing network traffic flow model and relationship between them, their proposed IDS is intended to automatically generates flow models and detect anomalies in network traffic that violate the established traffic flow model. The generation of all possible network is very difficult task for typical IP network due to diversity of devices, protocols and other multiple factor but in case of network model for SCADA it is possible to generate all possible traffic flow in SCADA network because of limited number of devices, protocols and regular communication pattern. The implementation and results of proposed IDS [34] is presented in [35] along with the challenges in modeling SCADA traffic, they used traditional five tuple protocol number, source and destination IP addresses and port numbers to generate network traffic flow information. Their dataset is comprises of the data from two SCADA facilities that also have corporate network. The field network contains Programmable Logic Control (PLC) and Remote Terminal Unit (RTU) devices. They used invariants for analyzing the SCADA network traffic. These invariants includes diurnal pattern of activity which describe network traffic based on time and weekdays, Log-normal connection sizes and Heavy-tail distributions which describe connection size distribution and self-similarity, in which whole message resemble the parts of the message . Their finding concludes that SCADA network have regular patterns and periodic in nature and both lognormal and heavy dont provide good fit for the SCADA connection size. They concluded that existing traffic models cannot be easily applied to SCADA traffic.

An anomaly detection system is proposed in [36], which analyze payload to detect anomalies and able to detect exploit based unknown attacks. The proposed IDS is consisting of four modules, namely a network sensor, a feature extractor, a similarity processor and the anomaly detection component. The network sensor is consist of Bro IDS which capture transport layer packets and reassemble it, the feature extractor received reassemble data and extract byte sequence and map it in a multidimensional feature space. The similarity processor analyzes the byte sequences and find similarities between them, which is decide on the distance between two sequences. Once the similarities are found, the new incoming data is analyzed by Anomaly detection system, it tries to fit the captured data with established byte sequence model and dissimilarities are presented as anomalies. The experiments are done for Hyper Text Transfer Protocol (HTTP) and Remote Procedure Call (RPC), the accuracy rate for HTTP rate is 88 % while 92 % for RPC traffic with false positive rate of 0.2% for both protocols.

Signature Based

An Intrusion Detection System for DNP 3 protocol is proposed in [37] which uses Bro [bro2008homepage], a network analysis framework. The proposed system modified the Bro and consists of three main components a network parser, an event handler and a Policy script interpreter which has protocol validation policy. The network packet parser is responsible for decoding incoming packet to byte stream as per DNP3 protocol specification and semantic information for each event. The event handler act as an interface between DNP3 parser and pol-

icy script interpreter. Event handlers are specified for each type of data field of DNP3 protocol. The protocol validation policy (PVP) is responsible to look for protocol specification violations, DNP3 protocol introduce some strict dependencies between different fields. The protocol validation policy performs two types of validation, inter-packet validation and intra-packet validation. In inter-packet validation looks for anomalous communication patterns, for e.g. unmatched request/reply, the system can keep the history of states from parsed network packets. This type of validation helps in detecting DoS and replay attacks. The Intra-packet validation is validates the dependencies between different data fields to detect malformed packets which can causes DoS, for e.g. an anomalous packet can have mismatch of length field and actual length of the real pay load. A simulated SCADA test bed is implemented for experimental purposes; they evaluate each module of the proposed system. They have conducted robustness and throughput testing in experiments but dont test their system for malicious activity. The system is robust to malformed packets and throughput of the system stands at 9247 packets/second. The proposed schemes or solution for the security of industrial control systems in this section will be able to detect anomalies or intrusion if it violates protocol specifications, normal communication patterns, and resource utilization profiles. The proposed schemes will not able to detect if a normal command with malicious value is send to the controlled devices. The command will satisfy protocol specification, communication patterns and the resource utilization of the devices will also be not affected. Signature based IDS are

not able to detect new attacks because of non-availability of signatures.

2.3 Machine Learning

2.3.1 Support Vector Machine

Support Vector Machine (SVM) has recently been introduced as a new technique, which perform binary classification of the data. SVM is an implementation of Vapniks's [38]structural risk minimization (SRM) principle. It plots the training vectors in highly dimensional feature space and creates a decision boundary between the classes. The features near to this decision boundary are known as support vectors. SVM perform classification on linearly separable data, in case of data which cannot linearly separable SVM use kernel function. The kernel function maps the data into high dimensional feature space to linearize the data. The various kernel functions can be used, such as linear, polynomial or Gaussian.

2.3.2 K-Nearest Neighbor

K-nearest neighbor (kNN) is a type of non-parametric classifier and special case of instance based classifier [39]. The kNN is trained is with the labelled training objects and classification of test object is done by computing the approximate distances or similarity between test object and the training object. The distance determines the nearest neighbor to the test object and the class of the K-nearest neighbor is assigned to the test object. The common distance calculation methods

are Euclidean and Manhattan [40] as shown in the equations below.

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

$$d(x, y) = \sqrt{\sum_{k=1}^{n} |x_k - y_k|}$$

Beside these, other distance calculation methods can be used for specific problems. Euclidean distance method is used in the proposed ADS. The value of k is also an important parameter; it determines the number of neighbor to be included in the nearest neighbor list. Smaller value of k make the scheme sensitive to noise whereas larger value will include object from other classes. The optimal value of the k can be obtained by cross-validation.

2.3.3 C4.5 Decision Tree

Classification tree is a prediction mode in machine learning, and it is also called Decision tree. It is tree pattern graph similar to flow chart structure; any internal node is a test property, each branch represents test result, and final nodes of leaves represent distribution situation of various types. C4.5 algorithm is a type of a decision tree which summaries training data in the form of a decision tree. Decision tree algorithms have proved popular due to their robustness and execution speed. It uses training data to build a decision tree. C4.5 employs a greedy approach

that uses an information theoretic measure as its guide. An attribute for root of the tree is chosen that divides the training instances into subsets. . It uses the gain ratio criterion in selecting the attribute for the root of the tree. The gain ratio criterion selects, from among those attributes with an average-or-better gain, the attribute that maximizes the ratio of its gain divided by its entropy. If the entropy of the class labels in these subsets is less than the entropy of the class labels in the full training set, then information has been gained through splitting on the attribute. The entropy is calculated by a function. The algorithm is applied recursively to form sub-trees, terminating when a given subset contains instances of only one class.

2.3.4 Genetic Algorithm

Genetic Algorithm is a type of inductive learning strategy that is inspired by the process of natural evolution and selection. Genetic Algorithm was by John Holland and his aides in 1975. Genetic algorithm initiate with the randomly selected chromosome with an objective function or problem to solve. Each chromosome is evaluated for the fitness score with respect to the given problem. The fitness score of a chromosome is calculated by using an evaluation function. The chromosomes with highest fitness score are selected to produce offsprings. This production is done by using one of two main genetic operators namely crossover and mutation. In crossover a point is selected in parent gene structure and exchange the remaining segment of the parent to create new offspring. Mutation is done by changing

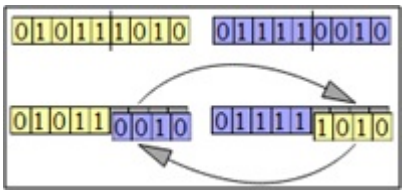

Figure 2.3: Gene Mutation

one or more component of a selected individual.

2.3.5 Best First Algorithm

Best rst search [41] is an AI search strategy that allows backtracking along the search path. Like greedy hill climbing, best rst moves through the search space by making local changes to the current feature subset. However, unlike hill climbing, if the path being explored begins to look less promising, the best rst search can back-track to a more promising previous subset and continue the search from there. Given enough time, a best rst search will explore the entire search space, so it is common to use a stopping criterion. Normally this involves limiting the number of fully expanded subsets that result in no improvement.

2.3.6 The Ripper System

The RIPPER System is probably the most popular technique representing this class. RIPPER [42]is a fast rule learning technique that generates concise rule sets. The system uses a set of rules and patterns that can prove realistic for classification for network traffic. The rule set generated by the system is simple and allows multiple rule sets to be created and used with a meta-classifier.

CHAPTER 3

DATASET, FEATURE SELECTION AND BASE-LINING

3.1 Introduction

In our research, we have used data of a turbine-generator operation from power generation plant. The purposed solution is a threshold based anomaly detection system. The threshold is based on the operational limits and average of a parameter. The dataset is evaluated using support vector machine (SVM), k-Nearest Neighbor (kNN) and C4.5 decision tree. In the following sections we describe each step of the experiment in details.

3.2 Feature Selection

The genetic algorithm is an efficient algorithm for feature selection where size of the feature set to save maximum information of the original dataset and the performance of the system is important. The goal of the feature subset selection is to identify and select a useful subset of features to in a larger set of often possibly irrelevant, mutually redundant, attributes with different associated measurement costs and/or risks. It can be presented as binary decision problem, in which each feature in the potential feature subset is considered as a binary gene and each individual consists of fixed-length binary string representing some subset of the given feature set. An individual of length x corresponds to a x-dimensional binary feature vector Y, where each bit represents the elimination or inclusion of the associated feature. It implies, x(i) = 0 represents elimination and x(i) = 1 indicates inclusion of the ith feature.

The main advantage of feature selection is improvement in the performance of the classification algorithm by reducing the complexity. This result in efficient and robust rules formation for classifying object into their respective class.

Each row in our dataset is comprised of 18 parameters related to temperature, pressure, electrical and safety indicators. The timing are very strict so it is not feasible to use all the parameters to identify anomalies. This is due to high incoming data rate and timely execution of all the processes in industrial applications. Feature selection is carried on dataset to reduce the number of features to differentiate normal and anomalous industrial operation. Genetic algorithm is used

to identify important attributes/features in the dataset. Features that affect the power output of the turbine-generator are selected using Genetic algorithm. The feature space reduced from eighteen to five important features which greatly affect power production, which are as follows,

i Fuel Gas Flow(Fuel GF): It is the amount of gas input to the combustion chamber. Its value varies as per the power requirement.

ii Main Shaft Vibration(MSV): It is the shaft vibration intensity. It is very important parameter, any great deviation from normal operation can persistently damage the turbine-generator system.

iii Gearbox Vibration(GBV): It is similar safety indicator as Main Shaft Vibration.

iv Exhaust gas temperature: It is temperature of turbine exhaust, which varies with the power output of the generator. It needs be monitored because excessive exhaust temperature can also damage the turbine-generator system.

v Several other parameters such Compressor discharge pressure, Lube oil supply Temperature directly affect power output of a turbine-generator system. Therefore, power is also selected to identify abnormal operations in this process.

3.3 Dataset

Each parameter of turbine-generator system should fall in pre-defined range to ensure normal system operation. The power requirement varies during different times of the day. The static nature of these processes can be used to model the normal operation behavior of the Gen-set.

The baseline parameters are established using the historical data from a power generation plant. The data set division is based on different power demand patterns during different time intervals. The dataset division is as follows:

i Group 1: 0600-1200 (Morning)

ii Group 2: 1200-1800 (Afternoon)

iii Group 3: 1800-2400 (Evening)

iv Group 4: 2400-0600 (Night)

The dataset was created for each group on the basis of attack sensitivity. Datasets are classified as follow,

i Highly Sensitive: A dataset row is labeled as an attack if any one of the selected parameter is compromised. This is labeled as X_1P, where X is group Number.

ii Medium Sensitive: A dataset row is labeled as an attack if any three of the selected parameters are compromised. This is labeled as X_3P, where X is group Number.

iii Least Sensitive: A dataset row is labeled as an attack if all of the selected parameters are compromised. This is labeled as X_5P, where X is group Number.

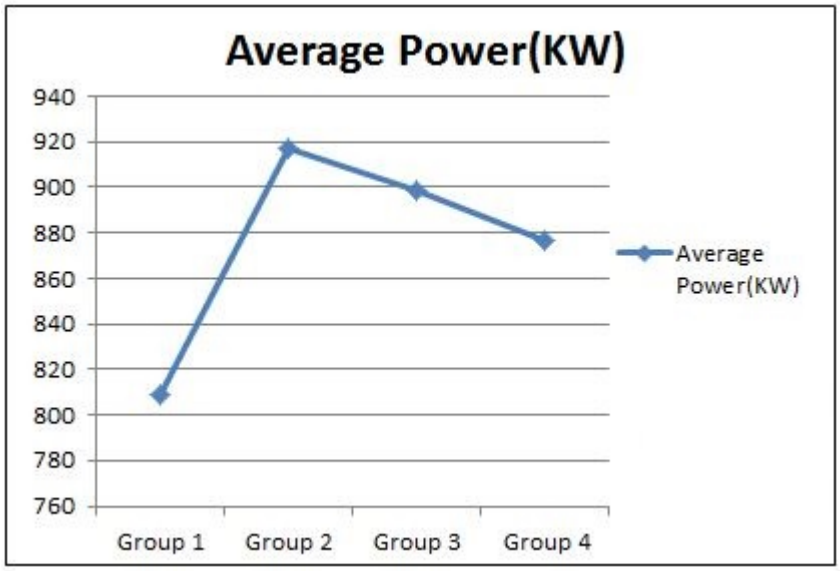

Figure 3.1: Average Power Demand

3.4 Testing Scenarios

To evaluate the performance of our proposed ADS, six testing scenarios are considered per dataset group. The difference between each testing scenarios is based on number of parameters considered for anomaly detection and total number of parameters in the dataset. These testing scenarios are as follow, 1. Highly Sensitive: A dataset row is labeled as an attack if any one of the parameter is compromised. 2. Medium Sensitive: A dataset row is labeled as an attack if any three of the parameter is compromised. 3. Least Sensitive: A dataset row is labeled as an attack all five of the parameter is compromised. The testing scenarios can be divided

into two classes namely reduced dataset and full dataset. The reduced dataset is consisting of the datasets with parameter that were selected from the full dataset using feature selection algorithms. The second class is consisting of datasets with all the parameter of the original dataset. This is done to evaluate the impact of the feature selection on the performance of the ADS. Three scenarios are selected for each class. These testing scenarios are shown in figure 3.2

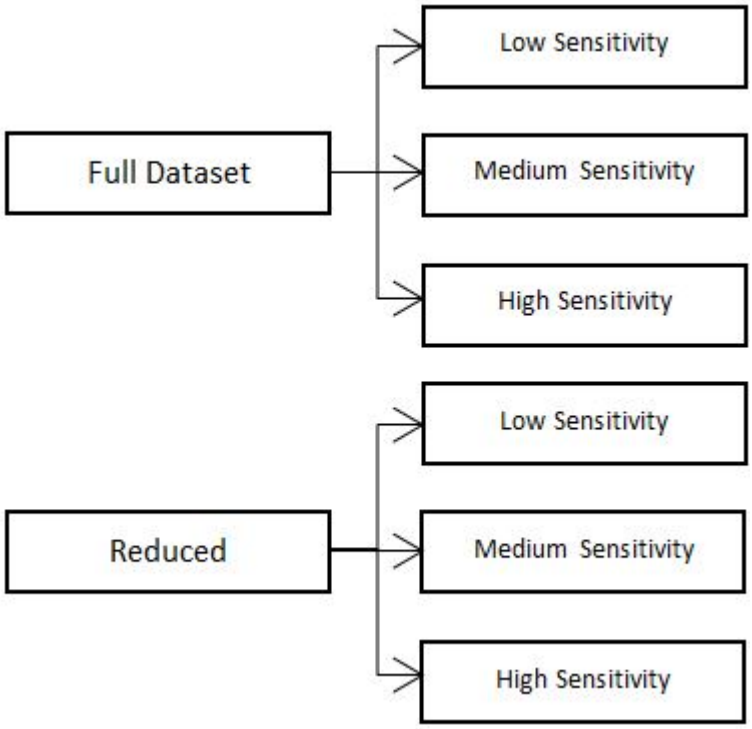

Figure 3.2: Testing Scenarios.

3.5 Base-Lining/Profiling

For each parameter maximum, minimum, trimmeans and thresholds values are calculated. Trimmeans were used instead of normal averages to exclude the outlier values, affecting the average. This helps in protecting the normal profile from an

attacker attempting to manipulate by sending periodic malicious values. It also protects the normal profile because of system faults.

An equation is developed for determining the threshold for each parameter as show in equation (3.1). The threshold is not static and incorporate the changing operational conditions.

$$T = ((\frac{OL - A}{OL}) * A) + A \tag{3.1}$$

where T is equal to Threshold for a parameter, OL is operational limit of parameter and A is trim-mean of the parameter. A parameter is considered to be compromised if its value exceeds the established threshold as per equation 3.1.

$$X = (\frac{OL - A}{OL}) \tag{3.2}$$

The percentage difference (X) of operational limit and trimmean is calculated using equation (3.2). This percentage is used to calculate tolerance region between trimmean and threshold. If the average value of a parameter is close to its operational limit, its tolerance region will be smaller as compared to a higher average value. For small tolerance region, small variation will result in an anomaly.

Sample rows of the dataset used and normal values of selected parameters are shown in Table 3.1, whereas Table 3.2 shows the anomalous values. These tables include operational limits of each parameter, computed thresholds, normal values and anomalous values. The anomalous values are shown in bold italics.

Table 3.1: Normal Operation

Parameter	Ex. Gas Temp	MST	GBV	Fuel GF	Power
OL	560	45	5	500	1120
Threshold	532	18.75	2.50	445	1032
	469	10.51	1.43	329	918
	416	11.62	1.56	319	834
	433	10.90	1.40	362	800
	458	12.42	1.60	371	819

Table 3.2: Anomalous Operation

Parameter	Ex. Gas Temp	MST	GBV	Fuel GF	Power
OL	560	45	5	500	1120
Threshold	532	18.75	2.50	445	1032
High Sensitive	470	12.636	1.365	***484***	884
Med. Sensitive	***547***	10.998	1.425	***474***	***1078***
Low Sensitive	***557***	***23.51***	***4.56***	***447***	***1103***

CHAPTER 4

ANOMALY DETECTION IN INDUSTRIAL CONTROL SYSTEM USING SUPPORT VECTOR MACHINE

4.1 Introduction

A Support Vector Machine (SVM) is a supervised machine learning system that has been widely employed for intrusion detection in the past. It is a binary classification algorithm that plots the training vectors in high dimensional feature space, separating the set of training vectors into two separate classes. The training samples close to a decision boundary constitute the support vectors. The SVM also enables users to balance between the number of misclassified samples and the

width of a decision boundary through a parameter called penalty factor. We utilize SVM in this research work due to its good generalization ability of the learning model. This implies that good accuracy can be achieved even with relatively smaller training datasets with SVM. Another advantage of SVM is its ability to handle a large number of features. In addition, SVM also ensures high accuracy for classification of future data from the same allotment to which the training data belongs. SVMs do not involve any reduction in the number of features and are free from the problem of over-fitting. Experimental results obtained confirm high accuracy of attack detection and insignificant false positive rates for the proposed approach.

4.2 Support Vector Machine

Support vector machines (SVMs)[43] belong to the class of supervised knowledge based systems that project the input vectors in feature space of large dimensions, assigning each vector a label. Data is categorized by SVMs by selecting a group of support vectors which are part of training inputs that outline a hyper plane in the feature space [38]. SVMs are known exhibit desirable results for both two-class and multi-class classification.

4.2.1 Two-Class Classification

For a two-class linearly separable data, SVM works by creating a hyperplane which divides the binary classes of the specified dataset with the largest margin. This

allows for the best generalization ability [43] Generalization ability is defined as the ability of the classifier to ensure classification accuracy for both training data and future data as well. The degree of separation between two classes is outlined with the help of a margin. Figure 4.1 below illustrates an optimal hyperplane for a two-class classification, where two different classes are represented: normal class (blue circles) and attack class (red circles).

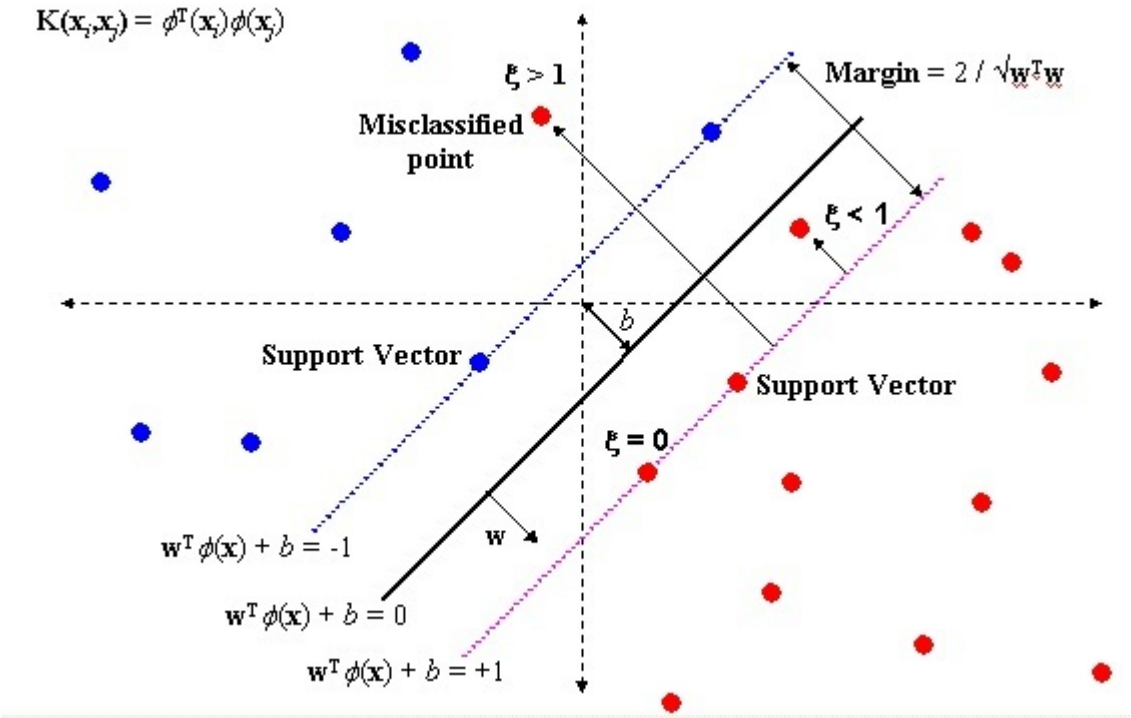

Figure 4.1: Support Vector Machine hyperplane for two input class

The objective of SVM here is to separate the two different classes by creating a linear boundary (solid line) that expands the margin (space between dashed lines) between the classes. As a result the data points that are found to be closest to the margin (circles on the dotted line) are called *support vectors*. The classifier is defined based on these support vectors. The model presented in Figure 4.1[44]

is valid for linearly separable data.

Mathematically, the linear boundary can be expressed as [45]:

$$w^T x + b = 0 \tag{4.1}$$

The classification problem using the training set can be estimated using a function $f : \mathbb{R}^n \mapsto \{\pm 1\}$. We present the normal class with $x \in normal, y = 1$ and the attack class with $x \in attack, y = -1$; $\{x_i, y_i\} \in \mathbb{R}^n X \{\pm 1\}$. If the training data can be linearly separated then there exists a pair $(w, b) \in \mathbb{R}^n X R$ such that

$$w^T x_i + b \geq +1 \;\; x_i \in normal \tag{4.2}$$

$$w^T x_i + b \leq -1 \;\; x_i \in anomalous \tag{4.3}$$

Therefore, the decision rule is given by

$$f_(w, b)x = sign(w^T x_i + b) \tag{4.4}$$

where w is the weight vector and b is the bias.

The optimal separating hyperplane that provide the largest margin between two classes can be found by minimizing squared norm of separating hyperplane.

$$[\frac{1}{2}||w||^2] \tag{4.5}$$

In cases where SVM is unable to separate the binary classes, a kernel function

is employed by SVM to plot the input vectors into n-dimensional feature spaces. The kernel function can be either linear, polynomial, or Gaussian. The number of parameters used by SVM is defined by the margin that splits the data points. This makes the SVMs free from the problem of over-fitting as they do not need any reduction in the number of feature. In addition, the prospect of generalization errors in SVMs is quite small. After classification is completed, an appropriate optimization process can be applied for identification of additional features if necessitated by the application [46].

4.2.2 Multi-Class Classification

Although, originally designed for binary classification, Support Vector Machines have recently been applied to the multi-class input space as well. Two main SVM schemes can be identified for classification of multi-class data, namely, one-against-all and one-against-one [47]. The one-against-all approach works by assembling and merging several binary classiers. On the other hand, one-against-one formulates all data into one optimization problem. Further details on multi-class classication can be obtained from [48].

4.3 Result and Analysis

The training of the algorithms is done on the 25% of the randomly selected data.The smaller training set is used because of less number of attack types. Similar approach is used in [31] where minimal training is done for an Intrusion

Detection System (IDS). Despite smaller training set, the proposed ADS able to perform efficiently. Each instance of the trained dataset is labelled as normal or anomalous. The labeling of the training dataset is done according to following criteria,

i *anomalous (1)*

ii *normal activity (0)*

The implementation of the SVM is done through Matlab. The total of 18 features are used to train SVM in all features based anomaly detection whereas five selected feature by feature selection algorithms are used to train SVM in reduced feature set based anomaly detection. The performance evaluation of the proposed system is done by using three important factors [49] these are as follows,

i *attack detection rate (ADR)*

ii *false positive rate (FPR)*

iii *system accuracy (Acc)*

The experiments tested attack detection rate, false positive rate and accuracy among full dataset and reduced feature input to SVM . The results of the experiment are presented for each type of sensitivity to attack i.e. High sensitive(1P), medium sensitivity (3P) and low sensitivity (5P). The experimental results for each group are presented below for each type of approach along with their averages.

High Sensitivity (1 Parameter)

The experimental results for the proposed approach with one parameter based anomaly detection are summarized in table 4.1. The achieved average accuracy with reduced dataset is 87.8% where as it is 71.3% for full dataset based approach. Performance degradation was observed for group 2 as the detection accuracy dropped to 80.6%. The minimum accuracy with full dataset based approach is 57.8% for group 4. This low detection accuracy is observed due to similarity between normal and anomalous data.

The average attack detection rate for the reduced data set based approach is 85%. The full dataset based approach achieved the maximum attack detection rate of 55.6% with the average detection rate of 46.7%. The average false positive rate with reduced dataset input is 11.3% whereas with the full dataset it is 20.6%. Group 4 showed the maximum false positive rate of 41.5% for full dataset.The precision rate for one parameter based approach with reduced dataset is approximately 90% while it is 80% for the full dataset based approach.

Table 4.1: SVM 1 Parameter Results

	Reduced Feature Set				Full Dataset			
	G1	G2	G3	G4	G1	G2	G3	G4
FPR	6.7	20.7	6.7	11.1	23.7	11.9	5.2	41.5
Acc	87.8	80.6	93.3	89.4	71.1	75.6	80.6	57.8
ADR	71.1	84.4	93.3	91.1	55.6	37.8	37.8	55.6
Precision	93.3	79.3	93.3	88.9	76.3	88.1	94.8	58.5

Medium Sensitivity (3 Parameter)

The experimental results improved greatly by increasing the number of parameters for anomaly detection. The ideal accuracy of 100% was archived for group 2 and group 3 with reduced dataset. The average accuracy is also very close to ideal. Another benefit is ideal false positive rate of zero across all groups. The attacks detection rate also improved from 85% to approx. 98%. These improved results are due to improved discrimination criteria between normal and anomalous data. Similar results are observed for full dataset approach. Its average accuracy increased to 92.1% from 71.3% but its attack detection rate still very low. The maximum achieved attack detection rate is 77.8% but average is 68.9%. Similar to reduce dataset approach the false positive rate is very close to ideal condition. The optimum precision rate is achieved with three parameter based approach for both full and reduced dataset based approach. In this approach, increasing the parameter increase the accuracy and false positive rate but the attack detection rate still not appropriate for real industrial control system. The poor attack detection rate is due to poor discriminating criteria which are due to and weak support vectors in hyperplane.

Table 4.2: SVM 3 Parameter Results

	Reduced Feature Set				Full Dataset			
	G1	G2	G3	G4	G1	G2	G3	G4
FPR	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0
Acc	98.9	100.0	100.0	98.9	89.4	94.4	93.3	91.1
ADR	95.6	100.0	100.0	95.6	57.8	77.8	75.6	64.4
Precision	100.0	100.0	100.0	100.0	100.0	100.0	99.3	100.0

Low Sensitivity (5 Parameter)

The ideal attack detection rate, accuracy and false positive rate were achieved across all groups with reduced data set as shown in Table 4.3. As the classes in the data set are either normal or anomalous, SVM is easily able to establish an optimal hyper-plane between normal and anomalous class.

Improved results were also observed with the full data set. The average accuracy was found to be 93.2% with minimal false positive rate across all groups. However, the attack detection rate with this approach is still not optimal. Despite improvement in some groups, the average attack detection shows insignificant improvement. The maximum attack detection rate is 84.4% for group 1 but the average attack detection rate is approx. 72.8%. The optimal precision rate for reduced dataset is maintained in this test case. However there is a decline in precision rate for full dataset based approach but the attack detection rate increases.

SVM Results Summary

The obtained results signify that increasing the number of parameters for anomaly detection improves the performance of the system. For all test scenarios the data set with reduced feature set was observed to outperform the full feature data set. The results from 5 parameter based approach exhibit ideal statistics but this approach is also known to be least sensitive to the attacks. On the other hand the one parameter based approach is highly sensitive to attacks but suffers from

low accuracy and attack detection rate. Its performance can also degrade due to noise which is quite common in industrial control system. The 3 Parameter based approach provides a balance between sensitivity to attack and performance characteristics. It has ideal false positive rate, high accuracy and attack detection rate. The comparison of accuracy, attack detection rate and false positive between reduced data set approach and full data set based approach is show in figures 4,5 and 6

Table 4.3: SVM 5 Parameter Results

	Reduced Feature Set				Full Dataset			
	G1	G2	G3	G4	G1	G2	G3	G4
FPR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Acc	100.0	100.0	100.0	100.0	96.1	93.9	91.7	91.1
ADR	100.0	100.0	100.0	100.0	84.4	75.6	66.7	64.4
Precision	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

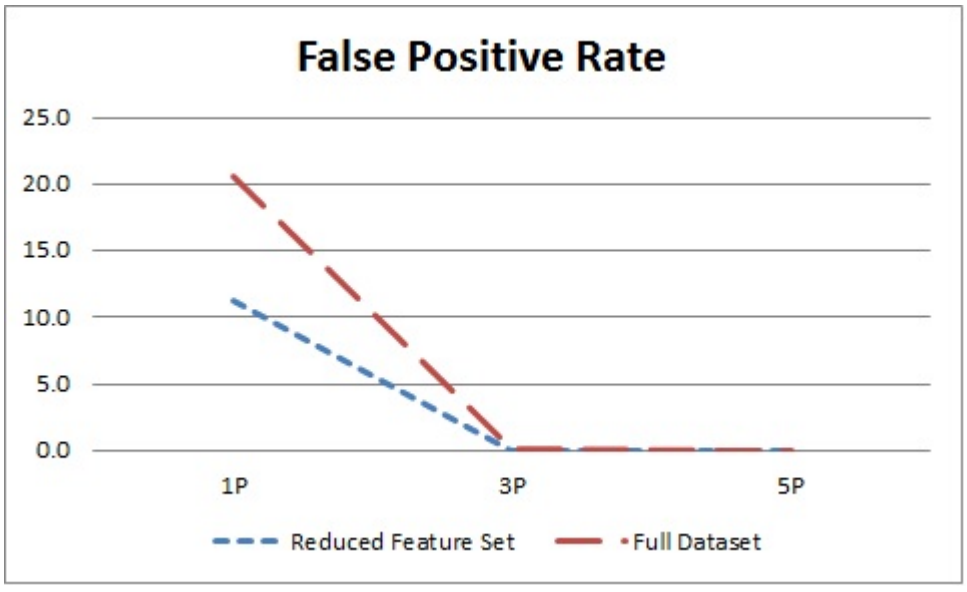

Figure 4.2: Average False Positive Rate for SVM

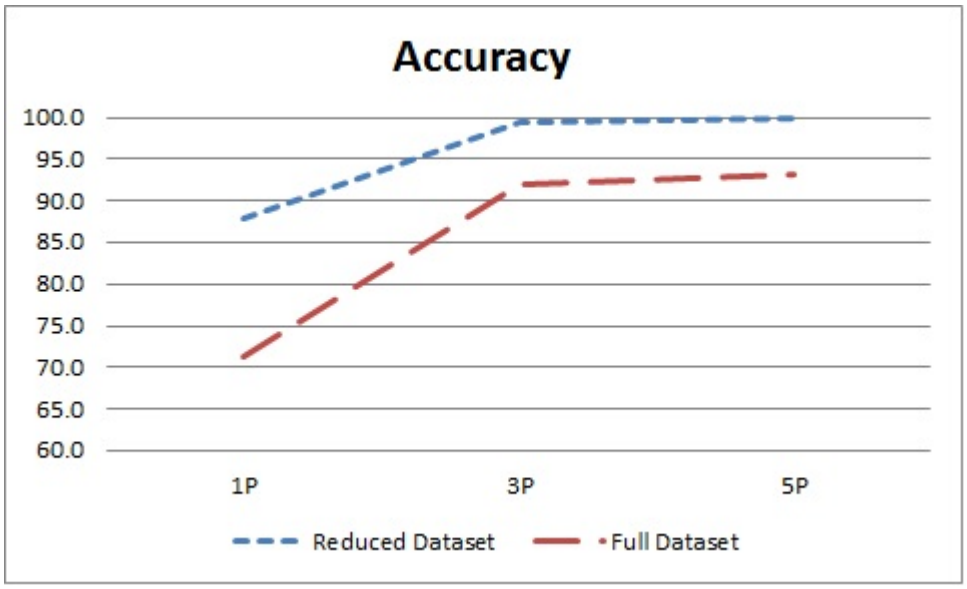

Figure 4.3: Average Accuracy for SVM

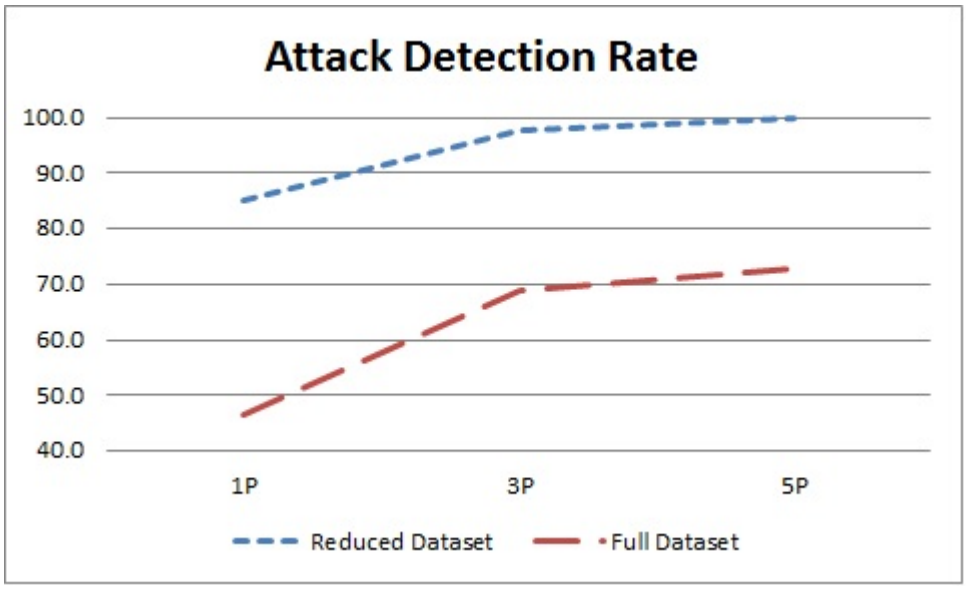

Figure 4.4: Average Attack Detection Rate for SVM

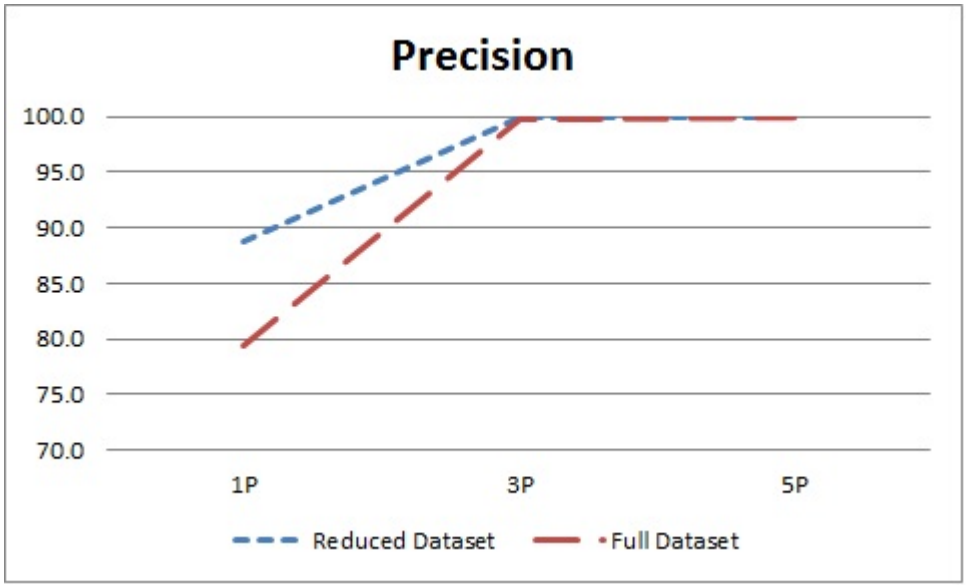

Figure 4.5: Average Precision for SVM

CHAPTER 5

ANOMALY DETECTION IN INDUSTRIAL CONTROL SYSTEM USING K-NEAREST NEIGHBOR

5.1 Introduction

The k nearest neighbor (kNN) algorithm is a well-known machine learning methodology characterized by good performance and a short training period. It is also commonly known as a lazy learning algorithm as learning does not initiate until the test sample is provided [1]. It is a complex algorithm capable of handling a large number of attributes and a large-scale dataset.

5.2 k-Nearest Neighbor

K nearest neighbor is an instance based learning algorithm. The instance based learning algorithms are lazy-learning algorithms because the generalization process does not start until classification is performed. Unlike the inductive learning approach, it does not contain the model training stage. This results in the less computational resources during the training phase but more computation is required during classification. The working principle of the k-Nearest neighbor algorithm based on the assumption that instance of the same class generally lie in close to each other. The label of the unknown instance can be determined by examining the class of the closest neighbors. In kNN the nearest neighbors to the query instance (unknown instance) are obtained by calculating the distance unknown instance and known instance. The objective is to find the single most frequent class label, the unknown instance is then labeled with frequent class label. The pseudo code of the kNN algorithm is illustrated below. The algorithm on how to compute the K-nearest neighbors is as follows,

- Determine the parameter K = number of nearest neighbors beforehand. This value is all up to you.
- Calculate the distance between the query-instance and all the training samples. You can use any distance algorithm.
- Sort the distances for all the training samples and determine the nearest neighbor based on the K-th minimum distance.

- Since this is supervised learning, get all the Categories of your training data for the sorted value which fall under K.
- Use the majority of nearest neighbors as the prediction value.

The instances are considered as points within an n-dimensional space, in which each direction corresponds to one of the n-features that describe an instance. The relative distance between unknown instance and known instance is calculated by using a distance metric. The objective of the distance metric is to minimize the distance between the two point of same class, while maximize the distance between different class. The most common distance formula are Euclidean and Manhattan as shown in equations 5.1 and equation 5.2 respectively,

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2]}$$

$$d(x, y) = \sqrt{\sum_{k=1}^{n} |x_k - y_k|}$$

However other distance metric are also used for special problems. To enhance the accuracy of the kNN, weighting scheme is used to influence the distance measurement and voting of each instance. In the kNN, the choice of k influences the performance of the algorithm. The kNN algorithm performance is also suscepti-

ble to the presence of the noise. If the unknown instance is located in the noisy region, the noisy instance will be in majority. This results in wrong labeling of the unknown instance, which degrades the performance of the algorithm. The effect of noise becomes insignificant with increasing value of the k. The experiment conducted in [1], concludes that kNN performance is not sensitive to noise when k was large. The small values of k are more robust on the most of datasets. However, 1NN out performs kNN on small datasets.

5.3 Result and Analysis

The training and testing criteria is same as described in SVM and it is also implemented in Matlab.

High Sensitivity (1 Parameter)

The results for each group based on the one parameter based approach are shown in Table 5.1. It can be confirmed from the results that kNN does not work well with either reduced data set based approach or full data set based approach. The attack detection rate is very low for both cases. For reduced data set the maximum attack detection rate is approx. 38%, which is obtained for group 3. The achieved average detection rate is 29%. The full data set based approach shows similar trend but the average attack detection rate is only 17%. In addition to low attack detection rate, this approach also has considerable high false positive rate. The average false positive rate is 15.5% for reduced data set based anomaly detection

where as it reaches to 28% when full data set is used. The full dataset based approach suffers from low precision rate of 71% while full dataset have precision rate of 85%.

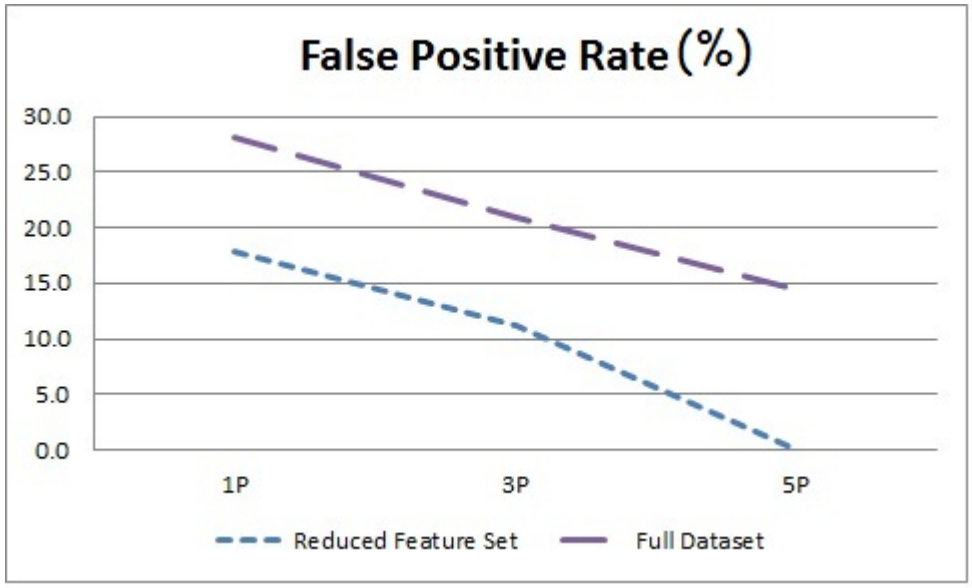

Figure 5.1: Average False Positive Rate for kNN

Consequently, the accuracy of both approaches was affected. The maximum accuracy for reduced data set is 75.5% for group 3. This group also has highest attack detection rate. The average accuracy is 70.5% for reduced data set whereas it is 58% for full data set.

Table 5.1: kNN 1 Parameter Results

	Reduced Feature Set				Full Dataset			
	G1	G2	G3	G4	G1	G2	G3	G4
FPR	14.1	18.5	11.9	17.8	32.6	23.0	34.1	23.0
Accuracy	70.6	69.4	75.6	66.7	55.6	61.1	52.8	62.8
ADR	24.4	33.3	37.8	20.0	20.0	13.3	13.3	20.0
Precision	85.9	81.5	88.1	82.2	67.4	77.0	65.9	77.0

Medium Sensitivity (3 Parameter)

Improved results are obtained for reduced features based approach when number of parameters for anomaly detection is increased to three. The maximum attack detection rate is increased to 77.8% whereas the average is improved to 74.4%. However, the attack detection rate does not meet the requirements of safety of real industrial control system. The full data set based approach still suffers from low attack detection rate. The average attack detection rate is 26.67% , which is only about 10% better than what obtained with 1 parameter. Similarly accuracy and false positive rate also show improvement using this approach. The minimum false positive rate achieved for reduced feature set is 3.7% but average is 9.8%. The average accuracy improves to 86.25%. In case of full data set based approach the average accuracy increased by 8% to 66%. The precision rate shows improvement for both full datastet and reduced dataset based approach. The average precision rate for reduced dataset based approach is 90% while the full dataset based approach achieved 80% of the precision rate.

Table 5.2: kNN 3 Parameter Results

	Reduced Feature Set				Full Dataset			
	G1	G2	G3	G4	G1	G2	G3	G4
FPR	16.3	3.7	8.1	11.1	20.0	22.2	25.2	16.3
Acc.	81.7	89.4	88.3	85.6	66.7	65.6	62.2	69.4
ADR	75.6	68.9	77.8	75.6	26.7	28.9	24.4	26.7
Precision	83.7	96.3	91.9	88.9	80.0	77.8	74.8	83.7

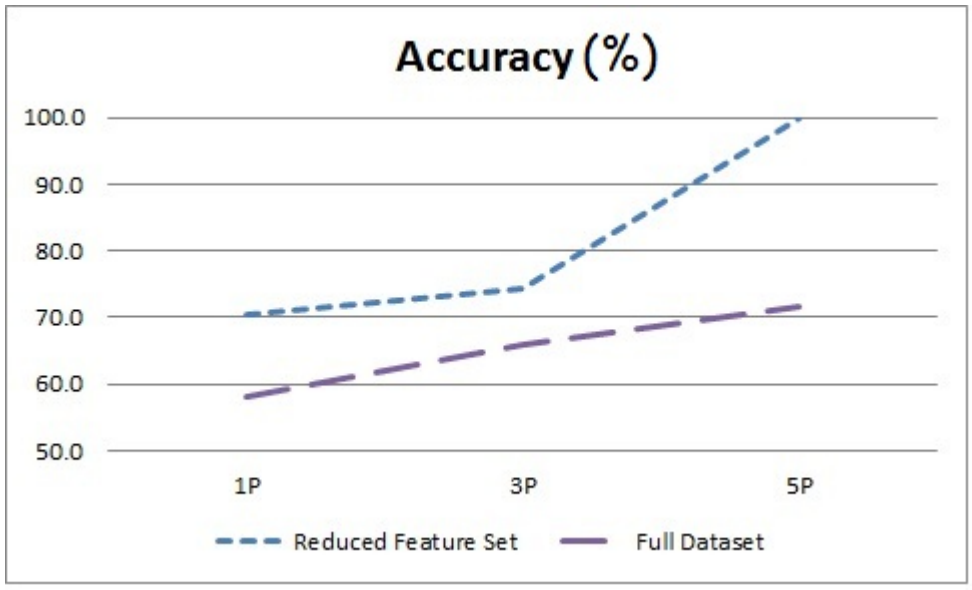

Figure 5.2: Average Accuracy for kNN

Low Sensitivity (5 Parameter)

Similar to SVM, the proposed approach with feature reduction produces ideal accuracy, false positive, attack detection rate and precision. The distance between a test sample and its neighbors will be calculated such that the test sample is always assigned the class of the training sample that is at minimum distance from it. As all training samples are either categorized as normal or anomalous, the distance between the normal and anomalous is always large.

The full data set based approach for 5 parameters did not show considerable improvement in attack detection rate. However, the maximum attack detection rate improved by only 2% whereas the average is 29.4%. In this case an attack row consists of 18 parameters with five of them in anomalous region. The distance calculated between the testing sample and the training data will also consider the 13 normal features. The incorrect nearest neighbor is found in this result. This

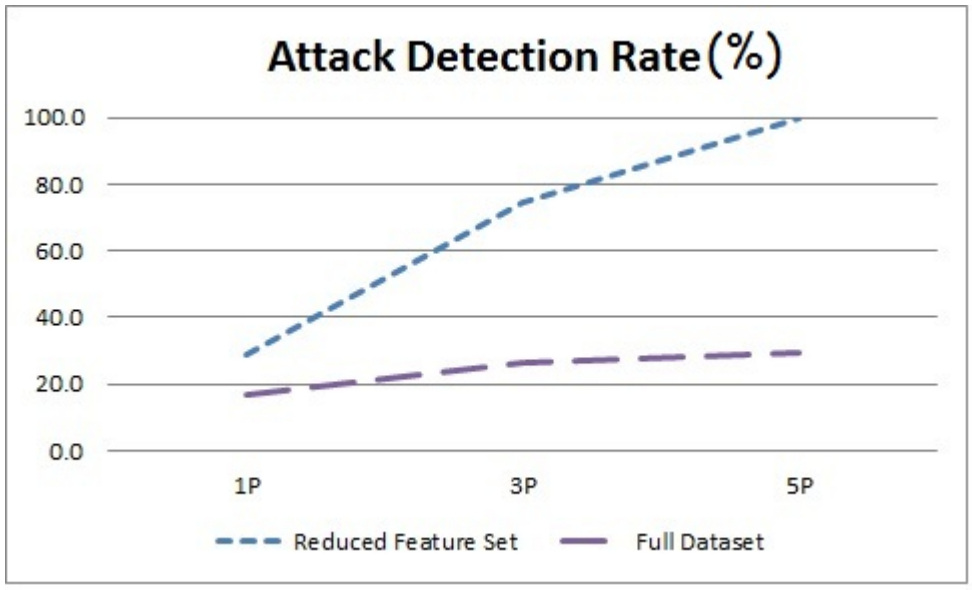

Figure 5.3: Average Attack Detection Rate for kNN

is due to the decrease in the false positive rate. The average false positive rate is 14.4% and the average accuracy improves to 71.5%. The average precision in this case for full dataset based approach is 85%.

Table 5.3: kNN 5 Parameter Results

	Reduced Feature Set				Full Dataset			
	G1	G2	G3	G4	G1	G2	G3	G4
FPR	0.0	0.0	0.0	0.0	12.6	7.4	23.7	14.1
Acc.	100.0	100.0	100.0	100.0	73.3	77.2	63.9	71.7
ADR	100.0	100.0	100.0	100.0	31.1	31.1	26.7	28.9
Precision	100.0	100.0	100.0	100.0	87.4	92.6	76.3	85.9

kNN Results Summary

The results obtained for kNN are similar to those obtained with SVM. Increasing the number of parameters for kNN also lead to improved results. The average attack detection rate, false positive rate and accuracy are shown in figure 6,7 and 8. However, the one parameter based approach did not produce acceptable

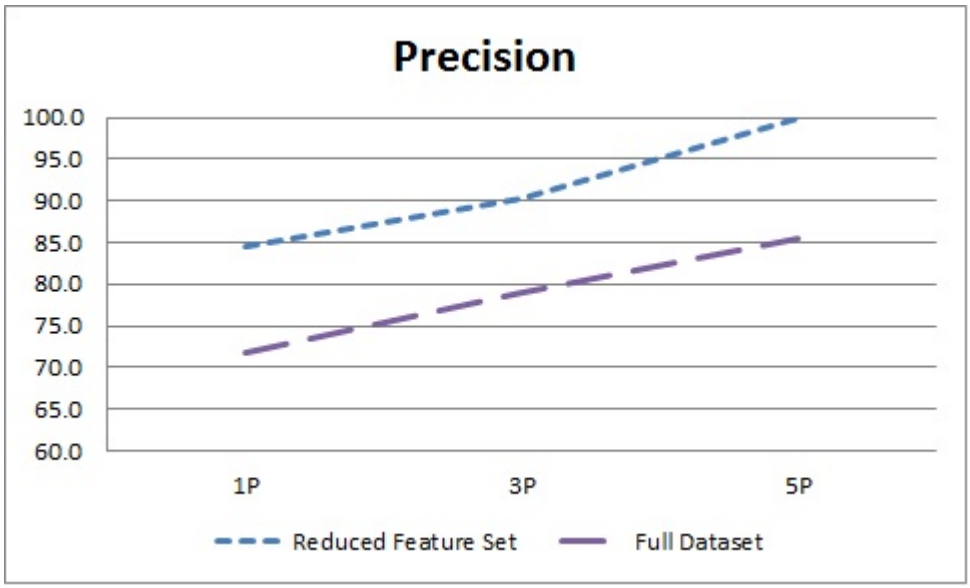

Figure 5.4: Average Precision for kNN

results with either reduced data set based approach or full data set based approach. Improved results are observed with the three parameter based approach but little improvement is evident with the full data set. The attack detection rate for this approach continues to decline further even with five parameter based approach. The attack detection rate for the reduced data set based approach show high improvement but may be not appropriate for industrial application because of average attack detection rate of 74.4%. The ideal results are achieved with reduced data set with five parameter based approach.

CHAPTER 6

ANOMALY DETECTION IN INDUSTRIAL CONTROL SYSTEM USING C4.5 DECISION TREE

6.1 Introduction

C4.5 is the most popular tree classifier that was introduced by Quinlan [50]. It is a descendant algorithm of CLS [51]and ID3 [52]. It forms a decision tree to describe the classifier and uses the decision tree to produce the rule set which is a group of if-then statements. C4.5 generates an original decision tree based on divide-and-conquer strategy described below.

Input: A training set S, a node T;

Output: A decision tree with the root T;

1. If the instance is S belong to the same class or the amount of instance is two few, set T as leaf node and label the node T with the most frequent class in S;

2. Otherwise,choose a test attribute X with two or more outcomes based on selection criterion, and label the node T with X;

3. Partition S into subsets $S_1, S_2, ...S_n$ according to the outcomes of attribute X for each instance; generate T'sn children nodes $T_1, T_2, ...T_n$;

4. For every group (S_i, S_i), build recursively a sub tree with the root T_i

6.2 Decision Trees

Information gain and default gain ratio are the two heuristics commonly used by C4.5 to rank possible tests [4]. While information gain is able to minimize the total entropy of the subsets, it shows heavy bias towards multi-valued attributes. Alternatively, the default option is gain ratio that divides information gain by the information provided by the test outcomes. Post the construction of decision tree, C4.5 needs to deal with the problems of discretization of numerical attributes, handling of missing value and pruning for the decision tree. The format of the test outcomes depends on the attributes of a dataset that can be either numeric or nominal. For a numerical attribute A, its output format is $A \geq h, A < h$, where the threshold h is found by sorting S on the values of A and

choosing the split between successive values that maximizes the criterion above. By default, an attribute A with discrete values has one outcome for each value, but there are options that permit combining at least two subsets as an output. To overcome the over-fitting problem of the training set, C4.5 performs pruning of the primary decision tree using post-pruning methods. A pessimistic statistical reasoning method by estimating the error rate forms the basis of the pruning algorithm. Assuming, N is the amount of instances in the instance set of the associated tree node, E is the number of errors, q is the true error rate, and f is the observed error rate. With a confidence threshold c (0.25 is the default value for C4.5); confidence boundaries z can be obtained according to equation 6.1

$$\Pr\left[\frac{f-q}{\sqrt{q\left(1-q\right)N}} > z\right] = c \tag{6.1}$$

The pessimistic error estimation of a tree node is gained according to the equation 6.2

$$e = \frac{f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} \tag{6.2}$$

C4.5 prunes from the leaves to the root. For all leaves in a sub tree, it computes their combination error estimation. If a leaf node replaces the sub tree, then C4.5 computes the error estimation of the leaf node. The leaf node substitutes the sub tree if the current error estimation is lower than the combination error estimation of the sub tree. The process repeats until no further replacements can take place.

6.2.1 Rule Set Classifiers

Due to the distributed nature of information about a single class in a tree, it may be difficult to interpret complex decision trees. As such, C4.5 uses groups for if-then rules for each class to classify a case. The first rule that matches the conditions outlined by a case classifies that case. In case no matching rule occurs, C4.5 assigns a default class to the case. The initial (unpruned) decision tree forms the C4.5 rule sets. Each path from the root of the tree to a leaf becomes a prototype rule whose conditions are the outcomes along the path and whose class is the label of the leaf. To simplify the rule the consequence of discarding each condition is studied. Dropping a condition may result in rise in the number N of cases covered by the rule, and the number E of cases that do not belong to the class nominated by the rule, and may lower the pessimistic error rate determined as above. The optimal pessimistic error rate is found by dropping conditions using a hill-climbing algorithm. To finalize the process, C4.5 picks a subset of simplified rules for each class and sorts them to minimize the error on the training cases. In addition, it chooses a default class. As a result, the obtained rule set usually has far fewer rules than the number of leaves on the pruned decision tree.

6.3 Result and Analysis

The C4.5 algorithm is executed in two phases: training phase and testing phase. The 25% of the dataset is used in the training of C4.5 and testing criteria is done on the 75% of the dataset.

6.3.1 Implementation

We have used Weka [53], which is an open source tool with Java implementation of data mining, machine learning algorithms, including data pre-processing, grouping, and association rule extraction. Weka is available under GNU general public license. It provides a graphical user interface for interaction. It has special file format Attribute-Relation File Format ".ARFF", in which data object are defined fixed number of attributes.

High Sensitivity (1 Parameter)

The experimental results showed for the proposed approach with one parameter based anomaly detection are summarized in table 6.1. The achieved average accuracy with reduced feature set is 92.4%. The average accuracy suffers due to low accuracy of 88.3% in group 2. The analysis of group 2 shows that the accuracy is impacted due to low high false negatives. This implies that wrong discriminating feature is selected in classification criteria. In case of full dataset the average achieved accuracy is 72.6%. Performance degradation was observed for group 1 as the detection accuracy dropped to 65%.

The average attack detection rate for the reduced feature based approach is 71.7%. The attack detection rate for group 2 is 53.3% which degrades the overall attack detection rate. The full dataset based approach achieved the maximum attack detection rate of 20.0% with the average detection rate of 15.0%. The

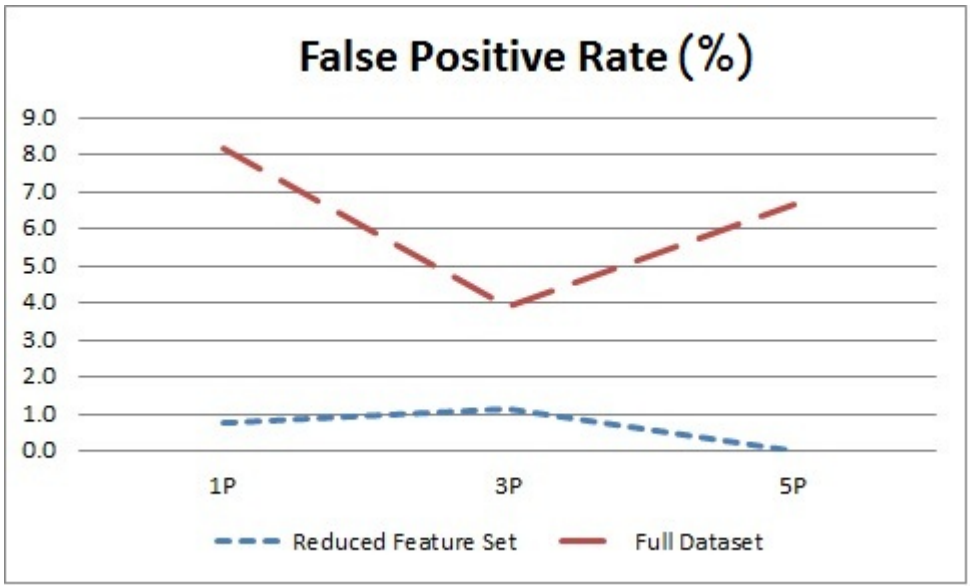

Figure 6.1: Average False Positive Rate for C4.5 Decision Tree

extremely low attack detection rate is because of wrong selection of feature in developing decision trees. The features that make a row anomalous can be any one of eighteen columns and it changes in every row. This results in selection of wrong discriminating feature.

The false positive rate is extremely low in all four groups in reduced feature set based approach. This implies that the accuracy of this approach is affected by number of false positives. The average false positive rate is 0.7%. In case of the full dataset based approach the average false positive rate is 8.1%. the

Table 6.1: C4.5 1 Parameter Results

2*	Reduced Feature Set				Full Dataset			
	G1	G2	G3	G4	G1	G2	G3	G4
FPR	1.5	0.0	0.0	1.5	18.5	0.0	2.2	11.9
Acc.	93.9	88.3	93.3	93.9	65.0	78.3	76.1	71.1
ADR	80.0	53.3	73.3	80.0	15.6	13.3	11.1	20.0
Precision	98.5	100	100	98.5	81.5	100	97.8	88.1

Medium Sensitivity (3 Parameter)

The results are considerably improved by increasing the number of parameter for anomaly detection to three. The attack detection rate of the full dataset based approach is improved from 15% to 51.7% but still not suitable for practical industrial control security solution. The improvement in the attack detection rate is reflected in the accuracy. The accuracy for full dataset based approach is improved to 85.0% from 72.6%. The false positive rate dropped to 3.9%

Similarly the results for reduced feature set based approach also improved. The average attack detection rate is 91.7% as compare to 1 parameter based approach. The accuracy is improved by 4.7% to 97.1%. The false positive rate is slightly increased from 0.7% to 1.1%. The ideal false positive rate is achieved except group 1. The false positive rate is increased from 1.5% to 4.4% but this increase is negligible. The results for three parameter based approach across four groups are shown in table 6.2.

Table 6.2: C4.5 3 Parameter Results

2*	Reduced Feature Set				Full Dataset			
	G1	G2	G3	G4	G1	G2	G3	G4
FPR	4.4	0.0	0.0	0.0	0.0	8.9	5.9	0.7
Acc.	96.7	96.7	97.8	97.2	88.9	79.4	82.2	89.4
ADR	100.0	86.7	91.1	88.9	55.6	44.4	46.7	60.0
Precision	95.5	100	100	100	100	91.1	94.1	99.3

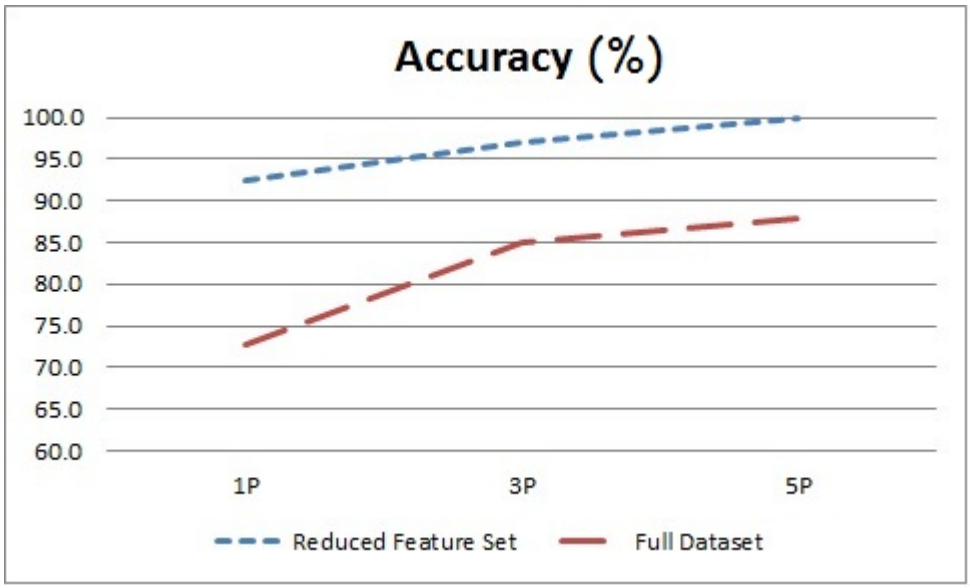

Figure 6.2: Average Accuracy for C4.5 Decision Tree

Low Sensitivity (5 Parameter)

The ideal results with reduced dataset are achieved with five parameter based approach, which is similar to SVM and kNN results. The analysis of the decision tree shows that the decision is based on the most discriminating feature. This results in the small tree size with binary decision logic.

Similar to other approach the best result for full dataset based approach is achieved with five parameters. The average detection rate improved by 19.4% to 71.1% with highest attack detection rate in group 4. The achieved average accuracy is also improved to 87.8%. However the false positive rate is increased to 6.7% when compare to three parameter based approach.

Summary C4.5 results follow the trend of SVM and kNN. The results improved by increasing the number of parameters for anomaly detection. The average false positive rate, accuracy, attack detection rate and precision rate is shown

in Fig 6.1, 6.2, 6.3 and 6.4 C4.5 achieved excellent precision rates with reduced data set for all test cases. The average precision rate is between 98% to 100% in all test cases The highest precision rate of 96% for full dataset based approach is achieved with the three parameter based approach. The precision rate fall to 93% for five parameter based approach. However the attack detection rate is increased in the five parameter based approach. The reduced features set based approach outperform full dataset based approach in all test cases. The ideal results are achieved with five parameter based approach but the three parameter based approach give the balance between performance and sensitivity to attacks. The three parameter based approach produced high accuracy and near ideal false positive rate. The attack detection rate is also acceptable for industrial control system security.

Table 6.3: C4.5 5 Parameter Results

2*	Reduced Feature Set				Full Dataset			
	G1	G2	G3	G4	G1	G2	G3	G4
FPR	0.0	0.0	0.0	0.0	9.6	0.0	3.7	13.3
Acc,	100.0	100.0	100.0	100.0	86.1	89.4	89.4	86.1
ADR	100.0	100.0	100.0	100.0	73.3	57.8	68.9	84.4
Precision	100.0	100.0	100.0	100.0	90.4	100.0	96.3	86.7

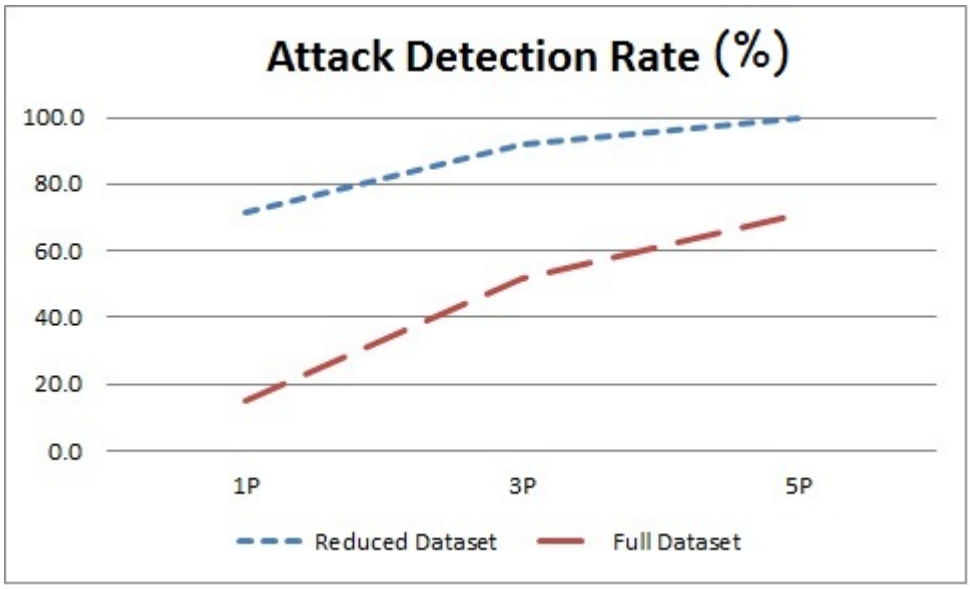

Figure 6.3: Average Attack Detection Rate for C4.5 Decision Tree

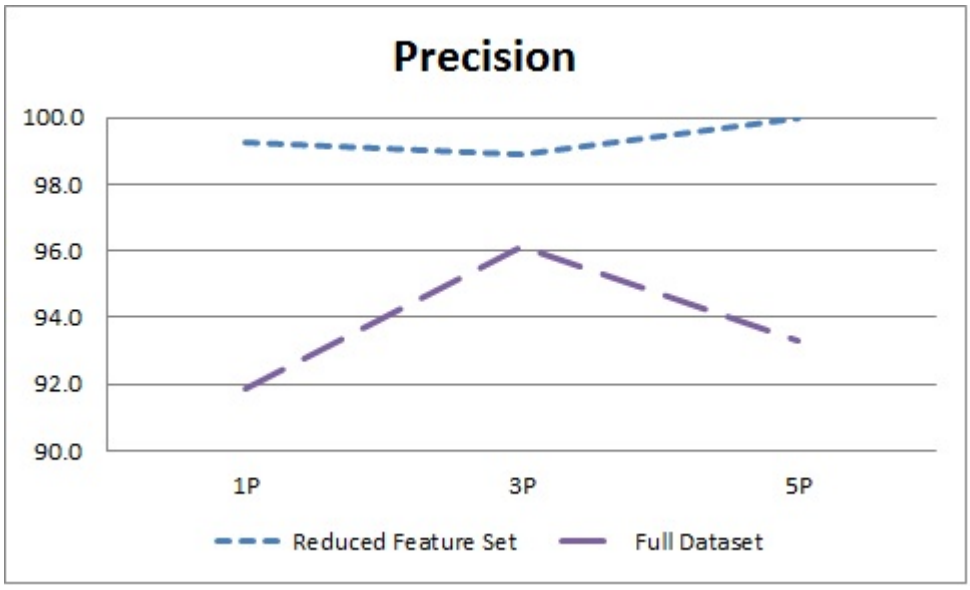

Figure 6.4: Average Precision for C4.5 Decision Tree

CHAPTER 7

CONCLUSION

The objective of this work was to develop an anomaly detection system for industrial control networks. The static nature of industrial control processes help to develop normal operation profiles that can be used to detect anomalous behavior and system faults. The proposed anomaly detection system uses feature selection to identify the relevant features to identify malicious behavior. It also helps in reducing the feature set for anomaly detection. The system uses Support Vector Machine, k-Nearest Neighbor and C4.5 algorithm. The proposed system shows that good attack detection rate and system accuracy. SVM and C4.5 produced accurate results even for high and medium sensitivity attacks. As compare to this kNN was unable to produce good result for low and medium sensitivity attacks test cases. The future work involves the testing of the proposed ADS in different industrial control system such as Oil and Gas industry. The future direction also involves in testing of different classification algorithm by incorporating physical system knowledge.

REFERENCES

[1] T. Turc and H. Grif, "Scada architecture for natural gas plant."

[2] A. Daneels and W. Salter, "What is scada," in *International Conference on Accelerator and Large Experimental Physics Control Systems*, 1999, pp. 339–343.

[3] M. Endi, Y. Elhalwagy, and A. Hashad, "Three-layer plc/scada system architecture in process automation and data monitoring," in *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, vol. 2. IEEE, 2010, pp. 774–779.

[4] N. TIB, "04-1, national communications system, technical information bulletin 04-1."

[5] IntelliSys, ""systemview"," http://www.intellisys-is.com/scada_software/systemview.aspx.

[6] Siemens, ""simatic step 7: The comprehensive engineering system"," http://www.automation.siemens.com/mcms/simatic-controller-software/en/step7/pages/default.aspx.

[7] B. Dutertre, "Formal modeling and analysis of the modbus protocol," in *Critical Infrastructure Protection*. Springer, 2007, pp. 189–204.

[8] "MODBUS.ORG", ""modbus over serial line specification implementation guide"," http://www.modbus.org/docs/Modbus_over_serial_line_V1.pdf.

[9] ""modbus messaging on tcp/ip implementation guide"," http://cars9.uchicago.edu/software/epics/Modbus_Messaging_Implementation_Guide_V1_0b.pdf.

[10] T. MicroWorks, "Dnp3 overview," *Raleigh, North Carolina (www. trianglemicroworks. com/documents/DNP3 Overview. pdf)*, 2002.

[11] K. Curtis, "A dnp3 protocol primer (revision a)."

[12] S. East, J. Butts, M. Papa, and S. Shenoi, "A taxonomy of attacks on the dnp3 protocol," in *Critical Infrastructure Protection III*. Springer, 2009, pp. 67–81.

[13] Finkle J. , "U.s. probes cyber-attack on water system," URL http://www.reuters.com/article/2011/11/19/cybersecurity-attack-idUSN1E7AH1QU20111119.

[14] Vijayan J. , "4 lessons from the springfield, ill. scada cyberattack," URL http://www.computerworld.com/s/article/9222113/4 lessons from the Springfield Ill. SCADA cyberattack23.

[15] GORMAN S. , "Electricity grid in u.s. penetrated by spies," URL-http://online.wsj.com/article/SB123914805204099085.html.

[16] G. A. Cagalaban, Y. So, and S. Kim, "Scada network insecurity: Securing critical infrastructures through scada security exploitation," *(Journal of Security Engineering)*, vol. 6, no. 6, p. 12, 2009.

[17] M. M. Fovino N. Igor, Coletta A, "Taxonomy of security solutions for the scada sector," ESCoRTS Consortium.

[18] E. P. Eric B., "Understanding vulnerabilities in scada and control systems," October 2004.

[19] P. Huitsing, R. Chandia, M. Papa, and S. Shenoi, "Attack taxonomies for the modbus protocols," *International Journal of Critical Infrastructure Protection*, vol. 1, pp. 37–44, 2008.

[20] A. A. Cárdenas, S. Amin, Z.-S. Lin, Y.-L. Huang, C.-Y. Huang, and S. Sastry, "Attacks against process control systems: risk assessment, detection, and response," in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*. ACM, 2011, pp. 355–366.

[21] R. Tsang, "Cyberthreats, vulnerabilities and attacks on scada networks," *University of California, Berkeley, Working Paper, http://gspp. berkeley. edu/iths/Tsang_SCADA% 20Attacks. pdf (as of Dec. 28, 2011)*, 2010.

[22] William T. Shaw, "Scada system vulnerabilities to cyber attack," URL http://www.electricenergyonline.com/?page=show articlemag=23article=181.

[23] J. Bigham, D. Gamez, and N. Lu, "Safeguarding scada systems with anomaly detection," in *Computer Network Security*. Springer, 2003, pp. 171–182.

[24] W. Gao, T. Morris, B. Reaves, and D. Richey, "On scada control system command and response injection and intrusion detection," in *eCrime Researchers Summit (eCrime), 2010*. IEEE, 2010, pp. 1–9.

[25] M. P. COUTINHO, G. LAMBERT-TORRES, L. EDUARDO, B. DA SILVA, and J. C. NETO, "Improving electric power system security against cybernetics attacks with intelligent techniques."

[26] Z. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.

[27] X. Jin, J. Bigham, J. Rodaway, D. Gamez, and C. Phillips, "Anomaly detection in electricity cyber infrastructures," in *Proceedings of the International Workshop on Complex Networks and Infrastructure Protection, CNIP*, 2006.

[28] M. Roesch *et al.*, "Snort-lightweight intrusion detection for networks," in *Proceedings of the 13th USENIX conference on System administration*. Seattle, Washington, 1999, pp. 229–238.

[29] I. Bro, "Homepage: http://www. bro-ids. org," 2008.

[30] S. Cheung, B. Dutertre, M. Fong, U. Lindqvist, K. Skinner, and A. Valdes, "Using model-based intrusion detection for scada networks," in *Proceedings of the SCADA Security Scientific Symposium*, 2007, pp. 1–12.

[31] N. Goldenberg and A. Wool, "Accurate modeling of modbus/tcp for intrusion detection in scada systems," 2013.

[32] D. Yang, A. Usynin, and J. W. Hines, "Anomaly-based intrusion detection for scada systems," in *5th Intl. Topical Meeting on Nuclear Plant Instrumentation, Control and Human Machine Interface Technologies (NPIC&HMIT 05)*, 2006, pp. 12–16.

[33] S. Parthasarathy and D. Kundur, "Bloom filter based intrusion detection for smart grid scada," in *Electrical & Computer Engineering (CCECE), 2012 25th IEEE Canadian Conference on*. IEEE, 2012, pp. 1–6.

[34] R. R. R. Barbosa and A. Pras, "Intrusion detection in scada networks," in *Mechanisms for Autonomous Management of Networks and Services*. Springer, 2010, pp. 163–166.

[35] R. R. Barbosa, R. Sadre, and A. Pras, "Difficulties in modeling scada traffic: a comparative analysis," in *Passive and Active Measurement*. Springer, 2012, pp. 126–135.

[36] P. Düssel, C. Gehl, P. Laskov, J.-U. Bußer, C. Störmann, and J. Kästner, "Cyber-critical infrastructure protection using real-time payload-

based anomaly detection," in *Critical Information Infrastructures Security*. Springer, 2010, pp. 85–97.

[37] H. Lin, A. Slagell, C. Di Martino, Z. Kalbarczyk, and R. K. Iyer, "Adapting bro into scada: Building a specification-based intrusion detection system for the dnp3 protocol," 2012.

[38] V. N. Vapnik, "An overview of statistical learning theory," *Neural Networks, IEEE Transactions on*, vol. 10, no. 5, pp. 988–999, 1999.

[39] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, no. 1, pp. 37–66, 1991.

[40] S. Salzberg. Pebls , "Pebls: Parallel exemplar-based learning system." http://www.cs.cmu.edu/afs/cs/project/airepository/ai/areas/learning/systems/pebls/0.html.

[41] R. E. Korf, "Linear-space best-first search," *Artificial Intelligence*, vol. 62, no. 1, pp. 41–78, 1993.

[42] W. W. Cohen and Y. Singer, "A simple, fast, and effective rule learner," in *Proceedings of the National Conference on Artificial Intelligence*. John Wiley & Sons Ltd, 1999, pp. 335–342.

[43] H. Xue, Q. Yang, and S. Chen, "Svm: Support vector machines," in *The Top Ten Algorithms in Data Mining*, 2009, ch. 3, pp. 37–59.

[44] "The standard svm formulation." [Online]. Available: http://research.microsoft.com/en-us/um/people/manik/projects/trade-off/svm.html

[45] X. Haijun, P. Fang, W. Ling, and L. Hongwei, "Ad hoc-based feature selection and support vector machine classifier for intrusion detection," in *IEEE International Conference on Grey Systems and Intelligent Services (GSIS 2007)*, Nanjing, China, Nov. 2007, pp. 1117–1121.

[46] T. Joachims, "Making large-scale support vector machine learning practical," in *Advances in Kernel Methods - Support Vector Learning*. Cambridge, USA: MIT Press, 1999, pp. 169–184.

[47] X. Bao, T. Xu, and H. Hou, "Network intrusion detection based on support vector machine," in *International Conference on Management and Service Science (MASS '09)*, Sep. 2009, pp. 1 –4.

[48] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415 –425, Mar. 2002.

[49] R.-C. Chen, K.-F. Cheng, and C.-F. Hsieh, "Using rough set and support vector machine for network intrusion detection," *arXiv preprint arXiv:1004.0567*, 2010.

[50] J. R. Quinlan, *C4. 5: programs for machine learning*. Morgan kaufmann, 1993, vol. 1.

[51] E. Hunt, J. Marin, and P. Stone, "Experiments in induction. 1966," 1986.

[52] J. R. Quinlan *et al.*, *Discovering rules by induction from large collections of examples.* Expert systems in the micro electronic age. Edinburgh University Press, 1979.

[53] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

Vitae

- Name: Muhammad Omer Qureshi
- Nationality: Pakistani
- Designation: Information Security Officer
- Date of Birth: 17/01/1989
- Email: *umerqureshi@outlook.com*
- Permenant Address: Karachi,Pakistan
- Under Graduate: Electronic Engineering, Sir Syed University of Engineering and Technology.
- Paper Submitted: Anomaly Detection in Industrial Control Networks.